2005

Risk properties of a Stein-like estimator for multinomial choice models

Vera Alexandrova Tabakova
Louisiana State University and Agricultural and Mechanical College, VTABAK1@LSU.EDU

RISK PROPERTIES OF A STEIN-LIKE ESTIMATOR FOR MULTINOMIAL CHOICE MODELS

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Economics

by
Vera Alexandrova Tabakova
M.A., Central European University, 1994
M.B.A., Bocconi University, 1996
M.S., Louisiana State University, 2000
May 2005

Acknowledgements

I would like to thank all my family and friends for supporting me and helping me complete the doctoral degree in economics.

Most of all I would like to thank my advisor Dr. R. Carter Hill, without whom this dissertation would have been impossible to complete. He not only gave me the knowledge necessary to pursue my work, but also guided me through all the steps of my academic career. He continuously inspired me, believed in me, and made me believe in myself.

I would like to thank the other committee members for their support and valuable comments: Dr. M. Dek Terrell, Dr. Douglas D. Schwalm, Dr. Eric T. Hillebrand, and Dr. Thomas C. Owen.

I thank all my friends for encouraging me and wanting me to succeed. I extend my special thanks to Anca and Ciprian for their continuous help and support, and for giving me time and peace of mind when I needed it most.

I thank my parents Ivana and Alexander, my 101-year-old grandmother Vera, and my sister Tatiana for motivating me to succeed by giving me their unconditional love and support, and being themselves a model to follow and aspire to.

I thank my husband Dan for his help, understanding and support through my long student career, and for the sacrifices he made to let me succeed. I thank our amazing son Alex for being a source of continuous joy and inspiration, and for reminding me why all my effort is worth it.

Table of Contents

Abstract

Stein-rule estimators, also known as shrinkage estimators, combine sample and non-sample information in a way that improves the precision of the estimation process or the quality of subsequent predictions. A Stein-rule estimator is a weighted average of a restricted and an unrestricted estimator, where the weights determine the degree of shrinkage, i.e. the importance that we place on the non-sample information. The existing literature shows that Stein-rule estimators may lead to squared error risk improvements in the linear regression, and in a number of non-linear models.

The dissertation explores Stein-rule estimation in the context of multinomial choice models. It consists of three main parts. First, a Monte Carlo study is conducted to examine the properties of a Stein-rule estimator for the orthonormal conditional logit model. The shrinkage estimator is compared to the maximum likelihood estimator based on different measures of risk, namely squared error risk, weighted error risk, risk of marginal effects, and mean squared error of prediction in-sample and out-of-sample. Secondly, the analysis is extended to a more general data generation process by introducing various degrees of collinearity within alternatives, or between alternative-specific variables. Finally, there are three applications of Stein-rule estimation in multinomial choice models using marketing data.

The main results of the study show that Stein-rule estimators offer significant risk improvement over the maximum likelihood estimator when certain conditions are met. The importance of this research is that shrinkage estimation is an easy to implement alternative to maximum likelihood estimation, which should be preferred in cases where we have good non-sample information, or when we are not sure of the performance of the MLE. The latter refers to data with small number of observations, or collinearity among the regressors, which is often a problem in practical applications.

1 Discrete Choice Models and the Use of Prior Information

1.1 Introduction

The purpose of this research is to provide comparisons of Stein rule and maximum likelihood estimation techniques in the context of the conditional logit model. Stein rule estimators, also known as shrinkage estimators, incorporate the use of non-sample information through a set of restrictions on the model parameters. Shrinkage reduces the absolute magnitude of the parameter estimates, which reduces their variability. As a result, shrinkage estimators have lower variance than the maximum likelihood estimator, but are generally more biased. We study the performance of the estimators based on their risk, which incorporates both the bias and the variance. Lower estimator risk implies that the obtained coefficients are closer to their true parameter values.

Section 1.2 discusses discrete choice models and the maximum likelihood estimation technique. Section 1.3 discusses different ways of introducing prior information into the estimation process. The Stein rule estimators are presented in Section 1.4.

1.2 Discrete Choice Models

Discrete choice models describe decision maker's choice among alternatives. These models are derived under the assumption of utility-maximizing behavior by the decision maker, and are also known as random utility models.

A decision maker i faces a choice among J different alternatives, and obtains a certain level of utility U_{ij} from each alternative. The utility is known by the decision maker, but not observable by the researcher. Some of the attributes of the alternatives,

and some of the characteristics of the decision maker are observable, and are used to model the expected utility $E\left[U_{ij}\right]$. Therefore, the utility of the i^{th} individual faced with J choices is

$$U_{ij} = E\left[U_{ij}\right] + \varepsilon_{ij} = x_{ij}'\beta + \varepsilon_{ij} \tag{1.1}$$

The residual ε_{ij} is the difference between the true utility U_{ij} and the part of utility captured in $E\left[U_{ij}\right]$.

The decision maker will choose the alternative that provides the highest utility. Therefore, alternative j will be selected if and only if $U_{ij} > U_{ik}$ for all $k \neq j$. The probability that alternative j yields the highest utility is

$$P_{ij} = \Pr\left[U_{ij} > U_{ik}\right] = \Pr\left[U_{ik} - U_{ij} < 0\right] \text{ , } k \neq j \tag{1.2}$$

The probability is a cumulative distribution with a density $f\left(\varepsilon_n\right)$. Different discrete choice models are obtained from choosing a specific distribution of unobserved utility.

- Conditional Logit Model

When the J disturbances are assumed to be independent and identically distributed with type I extreme value distribution, then the underlying model is *logit*. The key assumption is that the errors are independent, which means that the unobserved portion of utility for one alternative is unrelated to the unobserved portion of utility for another alternative. Logit is very popular because the formula for the choice probabilities has a closed form, which makes the model relatively easy to estimate. The probability that individual i chooses alternative j is:

$$P_{ij} = \Pr ob\left(Y_i = j\right) = \frac{\exp\left(x_{ij}'\beta\right)}{\sum_{j=1}^{J}\exp\left(x_{ij}'\beta\right)} \tag{1.3}$$

When the order of alternatives is not important and when the explanatory variables x_{ij} are individual and alternative specific, the model is called conditional logit. The most significant contribution to this model was done by Daniel McFadden (1974), who was awarded a Nobel prize for his work. The conditional logit has many applications in economics, marketing, transportation research, and other fields. Some of the popular examples in the literature analyze selecting mode of transportation, occupational choice, or choice among competing products, to name just a few.

- Maximum Likelihood Estimation

The unknown parameters of the conditional logit model can be obtained by maximum likelihood estimation. The probability function of the observable random variable y_i is $f\left(y_i\right) = P_{i1}^{\,y_{i1}} ... P_{iJ}^{\,y_{iJ}}$, where $y_{ij} = 1$ if individual *i* chooses alternative *j*, and zero otherwise. The likelihood function is the joint probability density function, which for a sample of *n* independent observations, is the product of the *n* probability functions.

$$L = \prod_{i=1}^{n} f\left(y_i\right) = \prod_{i=1}^{n} P_{i1}^{\,y_{i1}} ... P_{iJ}^{\,y_{iJ}} \tag{1.4}$$

Estimates of the parameters are obtained by maximizing the log-likelihood function, which is globally concave.

$$\ln L\left(\beta\right) = \sum_{i=1}^{n}\sum_{j=1}^{J} y_{ij} \ln\left(P_{ij}\right), \tag{1.5}$$

where y_{ij} equals to one if individual *i* chooses alternative *j*. Since the function is non-linear, estimation is usually done using the Newton-Raphson algorithm or the method of

scoring. The gradient and the Hessian are shown in equations (1.6) and (1.7) respectively.

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n}\sum_{j=1}^{J}\left(d_{ij} - P_{ij}\right)x_{ij} \tag{1.6}$$

$$\frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n}\sum_{j=1}^{J}\left(x_{ij} - \bar{x}_i\right)P_{ij}\left(x_{ij} - \bar{x}_i\right)', \text{ where } \bar{x}_i = \sum_{j=1}^{J} P_{ij}x_{ij} \tag{1.7}$$

- The Independence of Irrelevant Alternatives

One of the properties of the conditional logit model is called the independence of irrelevant alternatives (IIA), coming from the assumption that the errors are uncorrelated and that they have the same variance across categories. It implies that the odds of choosing one alternative versus another are the same no matter what other alternatives are available and what their characteristics are. The ratio of probabilities for alternative 1 and 2 is:

$$\frac{P_{i1}}{P_{i2}} = \frac{\exp\left(x_{i1}'\beta\right)\Big/\sum_{j=1}^{J}\exp\left(x_{ij}'\beta\right)}{\exp\left(x_{i2}'\beta\right)\Big/\sum_{j=1}^{J}\exp\left(x_{ij}'\beta\right)} = \frac{\exp\left(x_{i1}'\beta\right)}{\exp\left(x_{i2}'\beta\right)} \tag{1.8}$$

This ratio depends only on alternatives 1 and 2, so it does not depend on alternatives 3 to *J*, which are *irrelevant* alternatives to this choice. There are many situations in which the IIA property is a good representation of reality. In cases when the assumption does not hold, a more general model will provide better estimates of the unknown parameters. A class of models called generalized extreme value (GEV) models allows the unobserved portion if utility to be correlated over alternatives. These types of models are beyond the scope of our current research.

1.3 Use of Prior Information

Econometric models use data to estimate the values of unknown parameters. In many cases, however, the data are difficult to work with. In addition, the researcher may have non-sample information about the problem of interest. Such information is usually available from the underlying theory, previous empirical work, or experience with the data, and may include knowledge about the signs or values of the estimated coefficients. If the information is correct, it would seem useful to combine it with the sample information in subsequent analysis and statistical estimation.

In the classical framework, prior information may be introduced either by augmenting the sample information, through the likelihood function, or by modifying the parameter space. The latter is achieved through equality and inequality restrictions. In the case of exact restrictions, the new parameter space is of reduced dimensionality, which improves the precision of parameter estimates, because the available information is concentrated on a smaller set of parameters.

Exact prior information on linear combinations of the parameters can be expressed as

$$R\beta = r, \tag{1.9}$$

where R is a matrix of dimension J *x* K with $J \leq K$, β is a K *x 1* vector of unknown parameters, and it is assumed that R has rank J, which implies that the J equations do not contain redundant information about β. If we consider the classical linear regression model, the restricted estimator β^* is obtained by minimizing the sum of squared residuals, subject to the condition $r = R\beta$. If $\hat{\beta}$ is the OLS estimator then the restricted least squares estimator is

$$\beta^* = \hat{\beta} + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}\left(r - R\hat{\beta}\right) \tag{1.10}$$

Another way of introducing linear restrictions is by substitution through solving $R\beta = r$ for *J* of the parameters as

$$\beta_1 = R_1^{-1}\left[r - R_2\beta_2\right], \tag{1.11}$$

where $R = \left[R_1 \vdots R_2\right]$ is partitioned and R_1 is a *J x J* nonsingular matrix containing the restrictions; $X = \left[X_1 \vdots X_2\right]$, and $\beta' = \left(\beta_1', \beta_2'\right)$. Then β_1 may be substituted into $y = X_1\beta_1 + X_2\beta_2 + e$ to obtain

$$y = X_1\left[R_1^{-1}\left(r - R_2\beta_2\right)\right] + X_2\beta_2 + e,$$

which can be written as:

$$y^* = X^*\beta_2 + e, \tag{1.12}$$

where $y^* = \left(y - X_1R_1^{-1}r\right)$ and $X^* = \left(-X_1R_1^{-1}R_2 + X_2\right)$. Ordinary least squares estimates of β_2 obtained from equation (1.12) and the estimates of β_1 obtained from equation (1.11) are identical to the ones obtained from applying the restricted estimator β^* shown in equation (1.10).[1]

If the restrictions are correct, $R\beta = r$, it can be shown that β^* is unbiased. If the restrictions are not correct and $r - R\beta = \delta \neq 0$, then the restricted estimator is biased since

$$E\beta^* = \beta + (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}\delta \tag{1.13}$$

It can be shown that the difference between $\text{var}\left(\hat{\beta}\right)$ and $\text{var}\left(\beta^*\right)$ is a positive semidefinite matrix, which implies that the restricted least squares estimator is more

[1] See Judge et al (1988).

efficient than the OLS estimator. This result is true even if the imposed restrictions are not correct. Therefore, the decision whether to use the restricted estimator or not involves a tradeoff between bias and variance reduction. One way to evaluate this tradeoff is using a loss function. In general, a loss function $L\left(\beta,\tilde{\beta}\right)$, reflects the loss incurred by incorrectly guessing β using the estimator $\tilde{\beta}$. The quadratic loss function takes the form

$$L\left(\tilde{\beta},\beta;W\right)=\left(\tilde{\beta}-\beta\right)'W\left(\tilde{\beta}-\beta\right) \tag{1.14}$$

where W is a positive semidefinite weighting matrix. The loss function is random because $\tilde{\beta}$ is a random variable. Therefore, the expected loss, called the estimator's risk, is more appropriate in evaluating the estimator:

$$E\left[L\left(\tilde{\beta},\beta;W\right)\right]=\mathcal{R}\left(\tilde{\beta},\beta;W\right) \tag{1.15}$$

The most common choices of the weight matrix W are $W=I$, which defines the risk as the mean square error, and $W=X'X$, which defines the risk as the mean square error of in-sample prediction.

When two estimators are compared based on their risk functions, the preferred estimator is the one with smaller risk for β.

Figure 1.1 compares several estimators based on their risk functions[2]. For simplicity of presentation the squared error loss of the orthonormal model is chosen, where $X'X=I_K$, $R=I_K$, and $\delta=r-R\beta$ denotes the specification error in the prior restrictions. However, the following conclusions hold also for nonorthonormal regressors $\left(X'X\neq I_K\right)$ and general linear restrictions $\left(R\neq I_K\right)$. The risk is plotted against the extent to which restrictions are not met, which is a function of the specification error, and

[2] See Judge et al, p.87, Kennedy (98), p.197.

the unknown parameters β and σ^2. The risk of the unrestricted least squares estimator is equal to $\sigma^2 K$ because the estimator is unbiased and does not depend on the prior information. The risk of the restricted estimator is equal to the sum of squares of the restriction specification errors $\delta'\delta$ and is unbounded if the prior information is "bad", i.e. if the specification error is large. The two estimators have the same risk when $\delta'\delta = \sigma^2 K$ or when $\delta'\delta/2\sigma^2 = K/2$. Therefore, the restricted estimator dominates the unrestricted estimator if the prior information is "good", and the opposite is true when the specification error increases.

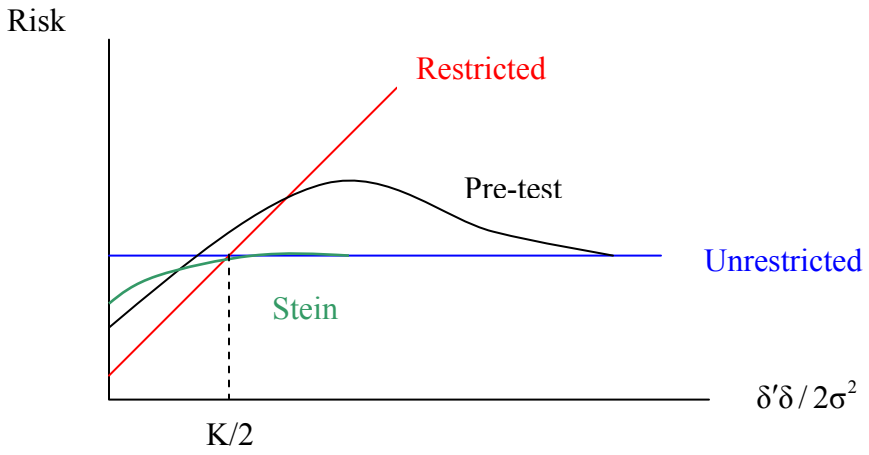


Figure 1.1: Risk functions for selected estimators

One way to address the choice of estimator is by testing the hypotheses that $H_0 : \delta = 0$ vs. $H_1 : \delta \neq 0$, or possibly $H_0 : \delta'\delta / 2\sigma^2 < (K/2)$ vs. $H_1 : \delta'\delta / 2\sigma^2 \geq (K/2)$. The first set of hypotheses tests if the imposed restrictions are correct; the second one is a test of whether the restricted estimator is superior to the unrestricted estimator. This

methodology defines what is called a pre-test estimator, because the choice of an estimator depends on the outcome of a preliminary test:

$$\tilde{\beta} = I_{[0,c)}(u)\beta^* + I_{[c,\infty)}(u)\hat{\beta}, \tag{1.16}$$

where $I_{[0,c)}(u)$ is an indicator function which takes the value one if the test statistic *u* falls in the interval [0,*c*), and zero otherwise, and *c* is the critical value of the F-distribution.

The pre-test estimator is superior to the estimator based on sample information only over a small portion of the parameter space. Moreover, the sampling properties of the pretest estimator are different from that of the restricted and unrestricted least squares estimators, and the usual standard errors associated with ordinary least squares or restricted least squares are no longer appropriate.

This result has lead to the development of non-traditional estimators that use non-sample information to modify the OLS estimator in such a way that the resulting estimator performs better over the entire parameter space, regardless of how correct the prior information is. Such estimators, originally proposed by Charles Stein in 1956, and are called Stein rule estimators.

1.4 Stein Rule Estimation

Stein rule estimators follow the work of Stein (1956) and James and Stein (1961) and combine sample information with non-sample information in a way that improves the precision of the estimation process and the quality of subsequent predictions. The Stein rule estimator uses a weighted average of the restricted and unrestricted estimators, the weight being a function of the magnitude of the F-statistic used to test the restrictions. In other words, it "shrinks" the unrestricted estimator towards the restricted estimator, and

the F-statistic determines the extent of shrinkage. Shrinkage estimators produce a biased estimator but may have lower estimation or prediction mean square error, or risk.

If $R\beta = r$ represent a set of $J \le K$ independent linear restrictions on β, the Stein rule estimator that combines sample and non-sample information is

$$\delta^*\left(\hat{\beta}, s\right) = \left[1 - \frac{as}{\left(r - R\hat{\beta}\right)'\left[R\left(X'X\right)^{-1}R'\right]^{-1}\left(r - R\hat{\beta}\right)}\right]\left(\hat{\beta} - \beta^*\right) + \beta^* \qquad (1.17)$$

where β^* is the restricted estimator described in equation (1.10). Sufficient condition for minimaxity, meaning that the estimator minimizes the maximum risk over the entire parameter space, are $J > 2$ restrictions and

$$0 < a < \frac{2}{T - K + 2}\left[\frac{tr\left\{\left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}W\left(X'X\right)^{-1}R'\right\}}{\eta_L} - 2\right] = a_{\max}, \quad (1.18)$$

where η_L is the largest characteristic root of the matrix in braces.[3] The estimator $\delta^*\left(\hat{\beta}, s\right)$ can be written as

$$\delta^*\left(\hat{\beta}, s\right) = \left(1 - \frac{c}{u}\right)\left(\hat{\beta} - \beta^*\right) + \beta^* = \left(1 - \frac{c}{u}\right)\hat{\beta} + \left(\frac{c}{u}\right)\beta^*, \qquad (1.19)$$

where u is the F statistic for the hypothesis $R\beta = r$,

$$u = \frac{\left(r - R\hat{\beta}\right)'\left[R\left(X'X\right)^{-1}R'\right]\left(r - R\hat{\beta}\right)/J}{s/\left(T - K\right)} \qquad (1.20)$$

and $c = a\left(T - K\right)/J$. If the data support the non-sample information then u will be small and a relatively large weight is placed on the restricted estimator β^*. Conversely, if the

[3] See Hill et al. (1991).

data do not support the imposed restrictions, u will be large and the unrestricted estimator $\hat{\beta}$ is more heavily weighted. If $u < c$ the Stein estimator reverses the sign of the estimator $\hat{\beta}$, or the latter is shrunk beyond the hypothesis vector. This problem is resolved by the use of the "positive rule" estimator $\delta^{+}\left(\hat{\beta}, s\right)$, which preserves the sign of the estimates and dominates the Stein rule estimator over the entire parameter space. It has the form

$$\delta^{+}\left(\hat{\beta}, s\right) = \left[1 - \frac{c}{u}\right]_{+}\left(\hat{\beta} - \beta^{*}\right) + \beta^{*}, \tag{1.21}$$

where $\left[a\right]_{+} = \max\left(a, 0\right)$. The weighting scheme of the Stein rule estimators insures that invalid prior information will not impose large losses in estimator efficiency relative to ordinary least squares. Some of the disadvantages of Stein rule estimator are that it is generally biased and requires the assumption of normally distributed errors. In addition, its covariance matrix depends on the unknown population parameters, so it cannot be used for hypothesis testing.

The dissertation is organized as follows. Chapter 2 discusses Stein rule estimation in nonlinear models and explores the empirical risk of the estimators in the context of the conditional logit model, where the explanatory variables are independent and have a standard normal distribution. Chapter 3 extends the analysis by allowing collinearity among the regressors. In both cases the analysis is conducted via a Monte Carlo simulation. Chapter 4 applies Stein rule estimation to three data sets in which we vary the quality of non-sample information. Chapter 5 summarizes our results.

2 Risk Properties of a Stein-Like Estimator for the Orthonormal Conditional Logit Model

2.1 Introduction

The Stein-rule estimator has lower risk than the maximum-likelihood estimator in the classical normal linear regression model if the number of hypothesis restrictions exceeds two and if certain other design related conditions are met[4]. Stein-like estimators may also lead to quadratic risk improvements in nonlinear regression models. In this paper we explore the risk properties of Stein-like estimation in the context of the conditional logit model. The analysis is done using a Monte Carlo simulation, in which the logit model explanatory variables are orthonormal. The choice of regressors is motivated by the fact that the original results for Stein-rule estimation, following Stein (1956) and James and Stein (1956), were derived for the orthonormal linear regression model. In the next chapter we extend the analysis to a more general design matrix, allowing multicollinearity among the regressors.

This chapter is organized as follows. In section 2 we present some results for shrinkage estimation in non-linear models. Section 3 describes the shrinkage estimators and the Monte Carlo design. Section 4 contains our empirical results, and section 5 concludes.

2.2 Stein Rule Estimation in Non-linear Models: Literature Review

There have been a number of studies on Stein-like estimation in the context of nonlinear models. Adkins and Hill (1989) use the approximate normality of MLE to construct a Stein-rule estimator for the probit model by replacing the elements of the

[4] Judge and Bock (1978)

Stein-rule used in the classical normal linear regression model with the estimates of the probit model. They find that when the sample size is small (50 observations), the Stein-like estimator outperforms the MLE in the sense that it has smaller risk over the range of parameters considered. For larger samples, the performance of all the estimators examined improves. The positive-part rule is superior to MLE and other Stein-rule alternatives for small to moderate degrees of hypothesis error.

Kim and Hill (1995) propose a positive-part Stein-like estimator for the Box-Cox model and derive the asymptotic risk functions of the maximum likelihood estimator, the restricted maximum likelihood estimator, the pretest estimator, and the positive-part rule under a sequence of local alternatives $H_0 : R\beta = r + \delta/\sqrt{T}$, where δ is a vector of constants defining the degree of hypothesis error. They adopt an asymptotic risk measure $AE\left[\left(\hat{\beta}-\beta\right)' \Im \left(\hat{\beta}-\beta\right)\right]$ where *AE* stands for asymptotic expectation, $\hat{\beta}$ is the estimator of β and $\Im$ is the information matrix. They show that under information weighted quadratic loss the risk of the shrinkage estimator for any $c > 0$ is smaller than the risk of the ML estimator, where *c* is a constant controlling the degree of shrinkage.

Adkins, Hill and Kim (1993) present different examples of nonlinear shrinkage estimation: linear model with autocorrelated errors, linear model with multiplicative heteroskedastic errors, several examples of the probit model, and two examples of nonlinear least squares. They find that in all cases but one, the Stein rules offer substantial risk gains over MLE and NLLS rules. The exception is the case of multiplicative heteroskedasticity where the non-sample information is very poor and the increase in risk over MLE is very small.

Adkins and Hill (1994) conduct a Monte Carlo study using a variety of economic models and sets of prior information, and conclude that in the context of probit model, using prior information via the Bayes, empirical Bayes or Stein estimator can significantly reduce estimator risk relative to pretesting or the MLE.

Following Adkins and Hill (1989) the Stein-like estimator for the logit model is

$$\delta = \left(1 - \frac{c}{u}\right)\beta_U + \left(\frac{c}{u}\right)\beta_R, \tag{2.1}$$

where β_U is the MLE (unrestricted) and β_R is the restricted MLE of β in the conditional logit model. In the absence of other non-sample information, we use the hypothesis restriction $H_0 : \beta = 0$. Under the null hypothesis, the restricted estimator β_R will be a $K \times 1$ vector of zeros. In the context of the conditional logit model this implies that all the alternatives are equally likely, i.e. if there are *J* alternatives, the probability of each alternative is 1/*J*. The shrinkage constant *c* and a test statistic *u* control how much unrestricted estimates are shrunk towards the restricted estimates. The test statistic can be based on a Lagrange multiplier (LM), likelihood ratio (LR), or Wald principles. Davidson and MacKinnon (1984) conduct a Monte Carlo experiment in the context of the probit model and conclude that the LR test performs better under the null hypothesis, and the LM test performs slightly better under the alternative hypothesis. Griffiths, Hill and Pope (1987) find that the distribution of the Wald statistics in small samples is a poor approximation of the asymptotic distribution. We use the likelihood ratio test statistic

$$LR = 2\left[\ln L(\beta_U) - \ln L\left(\beta_R\right)\right] \tag{2.2}$$

If the null hypothesis (*K* restrictions) is true, the likelihood ratio test statistic *LR* is asymptotically distributed as χ^2_K. If the restrictions agree with the data, the value of the

test statistic is relatively small, and the restricted estimator β_R is weighted more heavily. Otherwise, the value of *u* is relatively large and more weight is placed on the unrestricted estimator β_U. The degree of shrinkage is also controlled by the constant *c*. Judge, Hill and Bock (1990) motivate the Stein rule estimator by providing an empirical Bayes justification and suggest that the value of the shrinkage constant should be $g-2$, where *g* is the number of restrictions. Therefore, we need at least three restrictions in order to use Stein rule estimation. Other studies suggest different shrinkage constants and there is no truly optimal value of the shrinkage constant *c* in the sense that there is no dominant Stein rule estimator over the entire parameter space.

If the value of the shrinkage constant *c* exceeds the value of the test statistic *u*, the estimated coefficients change signs. To prevent this, the corresponding positive-part rule of the Stein rule estimator is given by

$$\delta^{+} = \left[1 - \frac{c}{u}\right]_{+} \beta_U + \left(\frac{c}{u}\right)\beta_R, \tag{2.3}$$

where $[\arg]_{+}$ is a function that chooses the maximum of the argument or zero, ensuring that δ^{+} is a convex combination or the restricted and the unrestricted maximum likelihood estimators. Shrinkage reduces the absolute magnitude of the parameter estimates, as compared to the unrestricted estimator, which reduces their variability. Therefore, shrinkage estimators have lower variance than the maximum likelihood estimator, but are generally biased. If the variance reduction is greater than the bias, the risk of the Stein rule estimators is lower than the risk of MLE.

2.3 Design of Monte Carlo Experiment

We have chosen four different Stein-rule estimators, and their positive counterparts, based on the degree of shrinkage. Following Judge et al. (1990) the base model uses a shrinkage constant $c = g - 2$. In addition, we choose a constant c_1, which shrinks the estimator less, and c_3 and c_4, which shrink the estimator more towards the restricted model. We change the shrinkage constant following Hill, Cartwright and Arbaugh (1991), who analyze the relative performance of several biased estimators using marketing data and find out that "overshrinking" can lead to significant risk improvement in out-of-sample prediction.

We name the Stein and Positive-part estimators "SteinXY" and "PsteinXY" respectively, where XY denotes alternative multiplicative factors of c. Specifically:

$$\begin{aligned} c_1 &= 0.5c: \text{ Stein05 and Pstein05} \\ c_2 &= c: \quad \text{Stein10 and Pstein10} \\ c_3 &= 1.5c: \text{ Stein15 and Pstein15} \\ c_4 &= 2c: \quad \text{Stein20 and Pstein20} \end{aligned} \tag{2.4}$$

In the Monte Carlo experiments we use 2000 Monte Carlo samples, generated for various degrees of specification error and other conditions as follows:

- The model we simulate has explanatory variables that are independent and have a standard normal distribution. Histograms of some marketing variables, such as prices, used in the context of conditional logit, motivate the choice of normal variables. Only one design matrix is generated per Monte Carlo experiment, and it remains fixed for each Monte Carlo sample.
- The "true" parameter vector β, used in the data generation process, is obtained as $\beta = w_i\beta_U + (1 - w_i)\beta_R$, where β_U is a Kx1 vector of ones, β_R is a

*K*x1 vector of zeros, and w_i = 0, 0.1, 0.2, … ,1 controls the degree of specification error. As w_i increases, β is further from β_R, meaning that the specification error in the restriction $\beta_R = 0$ is greater.

- For each Monte Carlo observation, the utility of individual *i* from alternative *j* is created as $U_{ij} = z'_{ij}\beta + e_{ij}$, where e_{ij} follows an extreme value (Gumbel) distribution. This assumption about the distribution of the unobserved portion of utility results in the logit model.[5] The observed variable y_{ij} is assigned a value of 1 if $U_{ij} = \max U_{ij}$, $j = 1,...,J$, and zero otherwise.
- For out-of-sample prediction, a holdout sample with N_O = 100 observations is generated as described above.
- In the Monte Carlo experiment, we manipulate the following:
    - number of variables: K = 4, 7, 10;
    - number of alternatives: J = 4, 7, 10;
    - sample size: N = 50, 250; The sample size of 50 is used to study the finite sample properties of the estimators. Such sample size was used in other Monte Carlo simulations to study the performance of Stein rule estimators.[6] The sample size of 250 is large enough to test the asymptotic properties of the estimators. We also performed Monte Carlo simulations with 500 observations but the results are not significantly different from the results with 250 observations and are not reported. After some experimentation we excluded sample sizes

[5] See Train (2003).
[6] See Adkins and Hill (1989).

smaller than 250 because the results did not provide any additional information.

- mean of the explanatory variables, in order to create three cases: (1) all alternatives have similar shares, (2) one alternative is dominant, (3) half of the alternatives are dominant;
- direction of the "true" parameter vector β, by setting the unrestricted vector $\beta_U = (-1,1,...,1)$. Changing the "true" β is necessary to explore the robustness of the estimation.

– For each estimator and each value of β the following estimates are obtained:

- goodness of fit measures and information criteria for model selection;
- squared and weighted risk for MLE and Stein rule estimators;
- risk of marginal effects;
- mean squared errors of prediction in-sample and out-of sample;
- hit rate in-sample and out-of-sample.

To compare the different models, we report two measures of goodness of fit, proposed by Estrella (1998), which are preferred in nonlinear models, and two information criteria. The goodness of fit measures are

$$E_1 = 1 - \left(\frac{LogL_U}{LogL_R} \right)^{-\left(\frac{2}{N}\right)LogL_R} \tag{2.5}$$

$$E_2 = 1 - \left(\frac{LogL_U - K}{LogL_R} \right)^{-\left(\frac{2}{N}\right)LogL_R} \tag{2.6}$$

where $LogL_U$ and $LogL_R$ are the values of the unrestricted and restricted log-likelihood functions, K is the number of parameters, and N is the number of observations. The two

measures are comparable to the R^2 in the linear regression model, with values ranging from 0 to 1, where 0 means no fit and 1 means perfect fit. E_2 corrects for the number of variables K and is preferred when comparing models where K varies. Unless stated otherwise, we refer to the latter when we report Estrella results, and we call it Estrella-R^2. In terms of model selection we use two information criteria: Akaike's information criterion (AIC) and Bayesian information criterion (BIC), where models with smaller AIC and BIC are preferred. Both criteria depend on the value of the log likelihood function, but penalize models with large number of parameters. The formulae and shown below.

$$AIC = -\frac{2}{N} LogL_U + \frac{2}{N} K \tag{2.7}$$

$$BIC = -\frac{2}{N} LogL_U + \frac{K}{N} Log\left(N\right) \tag{2.8}$$

In the Monte Carlo experiment, the risk performance of the estimators in equation (2.4) is examined under two loss functions. Squared error loss for estimator β_U is denoted

$$L_1 = \left(\beta_U - \beta\right)' \left(\beta_U - \beta\right) \tag{2.9}$$

and gives the squared distance between the unrestricted estimator β_U and the true parameter vector β. The weighted loss is computed as

$$L_2 = \left(\beta_U - \beta\right)' \left(\Im\right) \left(\beta_U - \beta\right), \tag{2.10}$$

where we estimate the information matrix $\Im$ using the negative of the Hessian.

Estimator asymptotic risk is average loss, and we use empirical risk to judge estimator performance. The variable of interest is the relative risk, i.e. the empirical risk

of each estimator divided by the risk of the MLE. The appendix contains the numerical results in terms of relative risk of the Stein rule estimators compared to the MLE. Values less that one indicate that the Stein estimators perform better than the MLE, and values greater than one show the reverse. We calculate a confidence interval for the risk of the MLE equal to two standard deviations from the estimated risk. If the lower bound of that confidence interval is greater than the risk of the Stein rule estimators, we consider the risk difference statistically significant. Shaded values of relative risk in the output tables indicate statistical significance.

Another measure of interest is the relative risk of the marginal effects. Marginal effects in conditional logit give the change in probability of choosing alternative j given a change in one of the explanatory variables, and can be calculated as

$$\frac{\partial P_{ij}}{\partial Z_{ik}} = -P_{ij} P_{ik} \beta \tag{2.11}$$

for $k \neq j$ and

$$\frac{\partial P_{ij}}{\partial Z_{ij}} = P_{ij} \left(1 - P_{ij}\right) \beta \tag{2.12}$$

The marginal effects depend not only on the parameter vector β but also on the probability of choosing alternative j and, therefore, on the values of the explanatory variables. In practice the estimated marginal effects may be more interesting than the estimated coefficients because they provide more relevant information to decision makers interested in influencing the choice of individuals. For example, in marketing context, the marginal effects help determine how is the probability of choosing a given brand going to be affected by a change in the price of that brand, and by a change in the price of competitive brands. The empirical risk is calculated as follows: first, we calculate the

"true" marginal effects by applying the above formulae to the true β. Next we calculate the estimated marginal effects using the MLE and Stein rule estimates. The loss for each Monte Carlo sample is calculated as the difference between the true and estimated marginal effects. The risk is the sum of the squared losses divided by the number of samples, and is computed for each level of specification error.

A very important aspect of this study is examining the performance of the estimators in terms of prediction in-sample and out-of-sample. We look at two different measures of prediction risk. First we calculate the mean squared error of prediction as

$$MSE_P = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\left(P_{ij} - \hat{P}_{ij}\right)^2}{N} \tag{2.13}$$

where P_{ij} is the true probability of individual i choosing alternative j, and $\hat{P}_{ij}$ is the predicted probability obtained using each of the nine estimators. The value we report is the average mean squared error of prediction over all Monte Carlo samples, and we do it in-sample and out-of-sample.

Following Kamakura and Wedel (2004), we also look at another measure of predictive accuracy:

$$MSE_Y = \frac{\sum_{i=1}^{N}\sum_{j=1}^{J}\left(Y_{ij} - \hat{P}_{ij}\right)^2}{N} \tag{2.14}$$

where Y equals 1 if an alternative was actually chosen, and zero otherwise. This measure was used in a Monte Carlo study to compare the performance of a finite mixture logit and mixed logit models in terms of out-of sample forecasting. The mean squared error is calculated as in the previous case, where we use y_{ij} instead of the true probability P_{ij}. The error is large if alternative j was actually chosen but we predict small probability of

selecting it, or if alternative *j* was not chosen but the predicted probability of selecting it is large. Intuitively, in practical applications, the probability of an alternative being chosen is less important than the actual choice of an alternative, which makes it logical to use this measure of predictive accuracy.

When performing the Monte Carlo experiments, we encountered the following problems. As we increased the number of variables *K*, we observed a jump in all risk values of the estimators for *N*=50 and large weights, usually for *w*=1. The values of the squared error loss and weighted error loss were more than 10 times larger at w=1 than the values of the risks for other weights. It was due to several Monte Carlo samples in which the estimated coefficients were 4-5 times larger than the true beta. These samples had very high goodness of fit values, usually over 98%. Likewise, when we increased the number of alternatives, we observed much higher risk as the true parameter vector increased in magnitude. The problem was more severe in the cases of one dominant or several dominant alternatives. The reason may be the existence of a perfect classifier[7]. The latter occurs when there exists some linear combination of the independent variables, say $X_{ij}'\widetilde{\beta}$, which allows us to predict the value of y_{ij} with perfect accuracy for every observation. In this case there is said to be complete separation of the data. As a result, we could not identify all the components of the parameter vector β. The existence of complete or quasi-complete separation of the data occurs in practice when the sample is very small, when almost all of the y_{ij}'s are equal to zero or almost all of them are equal to one, or when the model fits extremely well. This is why the problems in our simulation occurred for *N*=50 and models with one or more dominant alternatives, because the

[7] See Davidson and MacKinnon (2004), p.458.

probabilities of choosing these alternatives approach zero or one, which causes the estimation to break down. In the Monte Carlo simulation, we solved the problem by excluding from the analysis samples with values for Estrella-R^2 greater than 98%. When more than 10% of the Monte Carlo samples are replaced because of high values of Estrella-R^2, the results should not be considered in our analysis because they are not representative of our Monte Carlo experiment. We clearly indicate these cases in the text and in the output tables. In some cases, because of the long execution time of the programs, we removed the check for values of Estrella-R^2 and ran the original versions of the programs. For each of these cases we checked the weight at which the experiment breaks down and indicated it in the output tables. Finally, we observed a difference in the number of excluded samples when we changed the direction of the true β. Generally, less samples were excluded from the experiment when we used gam=$(-1,1,...,1)$ as the vector of unrestricted coefficients used to create β. Since the normal range of values of $x_{ij}'\beta$ should be $(-3,\ 3)$, having the same sign for all coefficients probably pushes the value towards one end of the interval and possibly out of the interval, while having different signs helps the values stay within the normal range.

2.4 Empirical Results

2.4.1 Four Equally Likely Alternatives

- Goodness of Fit

Table 2.1 shows the goodness of fit measures for each Monte Carlo experiment. The model fits the data very well when the signal-to-noise ratio increases. Estrella-R^2 reaches 0.8 when w=1. Towards the mid-range of our parameter space, Estrella-R^2 is over 40%, which is common in this type of models.

Table 2.1: Goodness of Fit Measures for Monte Carlo Designs, Orthonormal Model

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.05	0.03	0.14	0.28	0.41	0.52	0.61	0.69	0.75	0.80
			BIC	3.00	2.97	2.89	2.77	2.61	2.44	2.27	2.10	1.95	1.80	1.67
			LR	4.03	5.55	9.94	15.93	23.78	32.47	40.76	49.20	56.70	64.10	70.62
4	4	250	Estrella	-0.02	0.01	0.10	0.21	0.34	0.45	0.55	0.64	0.70	0.76	0.80
			BIC	2.84	2.81	2.73	2.60	2.45	2.29	2.13	1.98	1.84	1.72	1.61
			LR	4.00	11.57	33.11	65.04	103.81	143.24	182.80	220.63	254.12	285.90	313.80
4	7	50	Estrella	-0.14	-0.08	0.08	0.27	0.45	0.60	0.70	0.78	0.83	0.87	0.90
			BIC	3.17	3.12	2.95	2.72	2.48	2.24	2.03	1.85	1.69	1.56	1.45
			LR	7.32	10.09	18.62	29.77	41.80	53.85	64.41	73.45	81.32	87.94	93.31
4	7	250	Estrella	-0.03	0.03	0.16	0.32	0.47	0.59	0.68	0.75	0.81	0.85	0.88
			BIC	2.90	2.85	2.70	2.51	2.31	2.11	1.93	1.76	1.63	1.50	1.40
			LR	7.01	20.54	56.11	103.46	154.76	205.03	250.52	290.61	325.34	356.29	382.93
4	10	50	Estrella	-0.21	-0.13	0.06	0.27	0.44	0.58	0.67	0.74	0.80	0.84	0.87
			BIC	3.35	3.28	3.08	2.84	2.62	2.39	2.21	2.05	1.91	1.80	1.69
			LR	10.22	13.94	23.74	35.52	46.98	58.34	67.24	75.20	82.22	87.90	93.05
4	10	250	Estrella	-0.04	0.03	0.20	0.40	0.56	0.68	0.77	0.83	0.87	0.90	0.92
			BIC	2.95	2.88	2.69	2.45	2.20	1.97	1.78	1.61	1.47	1.35	1.24
			LR	10.15	28.02	74.62	135.79	197.48	255.00	304.42	346.08	381.69	411.78	437.88
4	4	50	Estrella	-0.08	-0.05	0.05	0.18	0.35	0.51	0.66	0.76	0.84	0.90	0.94
j_1 dominant			BIC	3.00	2.97	2.87	2.72	2.51	2.27	2.02	1.78	1.54	1.33	1.15
			LR	4.04	5.67	10.60	18.14	28.57	40.70	53.35	65.44	77.25	87.69	96.62
4	4	250	Estrella	-0.02	0.02	0.11	0.24	0.40	0.55	0.68	0.79	0.87	0.92	0.96
j_1 dominant			BIC	2.84	2.81	2.72	2.56	2.36	2.13	1.89	1.63	1.39	1.16	0.95
			LR	4.00	11.82	35.72	74.60	124.80	182.29	243.82	308.62	367.66	424.50	476.83
4	7	50	Estrella	-0.14	-0.08	0.10	0.32	0.51	0.67	0.77	0.85	0.90	0.93	0.95
j_1 dominant			BIC	3.17	3.11	2.93	2.67	2.39	2.11	1.86	1.64	1.45	1.30	1.19
			LR	7.32	10.29	19.58	32.38	46.49	60.69	72.85	83.91	93.51	100.89	106.38
4	7	250	Estrella*	-0.03	0.03	0.17	0.34	0.51	0.65	0.76	0.84	0.90	0.93	0.96
j_1 dominant			BIC	2.90	2.84	2.69	2.48	2.24	2.00	1.76	1.53	1.31	1.13	0.96
			LR	7.06	20.98	58.56	111.26	171.94	232.65	292.84	350.27	403.83	449.85	491.92
4	10	50	Estrella	-0.21	-0.13	0.08	0.33	0.53	0.70	0.81	0.88	0.92	0.95	0.96
j_1 dominant			BIC	3.35	3.27	3.06	2.77	2.47	2.15	1.87	1.63	1.45	1.31	1.22
			LR	10.22	14.12	24.79	39.12	54.09	70.02	84.13	96.16	105.14	112.04	116.76
4	10	250	Estrella	-0.04	0.03	0.22	0.44	0.64	0.77	0.87	0.92	0.96	0.98	0.99
j_1 dominant			BIC	2.95	2.88	2.67	2.38	2.06	1.76	1.48	1.23	1.02	0.84	0.69
			LR	10.15	28.90	80.72	153.00	232.29	309.36	379.09	441.66	494.10	539.15	576.40
4	4	50	Estrella	-0.08	-0.05	0.05	0.18	0.34	0.50	0.64	0.74	0.81	0.86	0.90
half dominant			BIC	3.00	2.97	2.87	2.73	2.52	2.29	2.06	1.84	1.64	1.47	1.32
			LR	4.04	5.69	10.63	18.00	28.09	39.55	51.49	62.53	72.43	80.77	88.17
4	4	250	Estrella	-0.02	0.02	0.11	0.24	0.40	0.55	0.67	0.77	0.84	0.89	0.93
half dominant			BIC	2.84	2.81	2.72	2.56	2.36	2.14	1.91	1.68	1.47	1.29	1.13
			LR	4.01	11.89	35.87	74.42	124.05	180.69	238.52	296.20	346.99	392.34	433.42
4	7	50	Estrella	-0.14	-0.08	0.10	0.31	0.51	0.66	0.77	0.84	0.89	0.92	0.94
half dominant			BIC	3.17	3.11	2.93	2.67	2.39	2.12	1.88	1.66	1.49	1.35	1.24
			LR	7.32	10.26	19.55	32.28	46.42	60.24	72.07	83.01	91.30	98.45	103.79
4	7	250	Estrella	-0.03	0.03	0.17	0.35	0.52	0.66	0.77	0.84	0.89	0.93	0.95
half dominant			BIC	2.90	2.84	2.69	2.47	2.22	1.97	1.73	1.53	1.34	1.18	1.05
			LR	7.00	21.12	59.90	114.44	176.55	239.52	298.48	350.42	396.79	436.94	470.22
4	10	50	Estrella	-0.21	-0.12	0.10	0.34	0.55	0.71	0.82	0.89	0.93	0.95	0.96
half dominant			BIC	3.35	3.27	3.05	2.76	2.43	2.12	1.84	1.61	1.43	1.32	1.25
			LR	10.22	14.25	25.42	39.88	56.01	71.67	85.96	97.37	106.44	111.71	115.25
4	10	250	Estrella	-0.04	0.03	0.22	0.44	0.62	0.76	0.85	0.91	0.94	0.96	0.97
half dominant			BIC	2.95	2.88	2.68	2.39	2.09	1.79	1.53	1.31	1.13	0.98	0.90
			LR	10.15	28.54	79.36	150.17	226.82	300.48	365.15	421.05	466.44	502.96	524.39

(Table continued)

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
7	4	50	Estrella	-0.08	-0.05	0.06	0.21	0.36	0.51	0.64	0.73	0.81	0.86	0.90
			BIC	4.12	4.09	3.98	3.81	3.60	3.37	3.12	2.89	2.67	2.46	2.28
			LR	3.99	5.85	11.44	19.75	30.11	41.57	54.23	65.76	76.98	87.01	96.18
7	4	250	Estrella	-0.02	0.02	0.11	0.25	0.39	0.52	0.63	0.72	0.79	0.84	0.87
			BIC	3.96	3.93	3.83	3.67	3.48	3.28	3.07	2.86	2.67	2.49	2.33
			LR	4.00	12.61	37.98	76.45	124.61	176.28	228.72	279.83	327.37	372.35	412.08
7	7	50	Estrella	-0.15	-0.08	0.08	0.28	0.47	0.61	0.73	0.80	0.86	0.89	0.92
			BIC	4.30	4.23	4.06	3.84	3.56	3.29	3.02	2.80	2.59	2.41	2.27
			LR	7.16	10.36	18.80	30.20	44.13	57.43	70.84	82.10	92.57	101.56	108.39
7	7	250	Estrella	-0.03	0.03	0.18	0.37	0.55	0.69	0.79	0.85	0.90	0.93	0.95
			BIC	4.02	3.96	3.79	3.55	3.26	2.97	2.70	2.46	2.25	2.07	1.92
			LR	7.01	22.03	64.02	124.99	195.83	268.09	335.42	396.51	448.40	493.73	532.71
7	10	50	Estrella	-0.21	-0.12	0.09	0.34	0.55	0.70	0.80	0.86	0.90	0.93	0.95
			BIC	4.47	4.38	4.17	3.86	3.53	3.21	2.92	2.68	2.46	2.30	2.18
			LR	10.21	14.60	25.25	40.81	57.22	73.26	87.64	99.92	110.50	118.67	124.51
7	10	250	Estrella	-0.04	0.04	0.25	0.47	0.65	0.78	0.86	0.91	0.94	0.96	0.97
			BIC	4.07	3.99	3.76	3.44	3.10	2.78	2.49	2.24	2.03	1.85	1.72
			LR	10.18	30.74	88.95	168.25	252.87	332.64	405.16	468.09	519.85	565.39	598.19
7	4	50	Estrella	-0.08	-0.04	0.08	0.27	0.49	0.68	0.82	0.91	0.95	0.97	0.97
j_1 dominant			BIC	4.12	4.08	3.95	3.73	3.41	3.02	2.60	2.21	1.90	1.73	1.67
			LR	4.00	6.01	12.64	23.95	39.72	59.09	80.19	99.81	115.32	123.70	126.93
7	4	250	Estrella*	-0.02	0.02	0.13	0.29	0.48	0.66	0.80	0.90	0.96	0.98	0.99
j_1 dominant			BIC	3.96	3.93	3.81	3.62	3.35	3.01	2.62	2.21	1.79	1.42	1.08
			LR	4.02	13.11	41.42	89.34	157.66	242.88	339.99	443.21	547.08	640.81	725.57
7	7	50	Estrella	-0.15	-0.08	0.10	0.32	0.55	0.73	0.85	0.92	0.95	0.97	0.97
j_1 dominant			BIC	4.30	4.23	4.05	3.78	3.41	3.02	2.63	2.25	1.99	1.85	1.79
			LR	7.17	10.41	19.56	33.15	51.67	70.91	90.49	109.72	122.33	129.53	132.65
7	7	250	Estrella*	-0.03	0.03	0.20	0.41	0.62	0.78	0.88	0.94	0.97	0.99	1.00
j_1 dominant			BIC	4.02	3.96	3.77	3.49	3.13	2.74	2.34	1.97	1.62	1.32	1.05
			LR	7.06	22.74	67.96	139.65	228.64	326.64	426.20	519.51	607.38	682.00	747.99
7	10	50	Estrella	-0.21	-0.11	0.12	0.40	0.64	0.79	0.88	0.93	0.96	0.97	0.97
j_1 dominant			BIC	4.47	4.38	4.14	3.77	3.35	2.95	2.58	2.27	2.07	1.95	1.89
			LR	10.21	14.78	26.85	44.97	66.07	86.21	104.91	120.07	130.15	136.11	139.04
7	10	250	Estrella*	-0.04	0.05	0.26	0.50	0.70	0.83	0.91	0.96	0.98	0.99	1.00
j_1 dominant			BIC	4.07	3.98	3.74	3.39	2.98	2.59	2.22	1.88	1.56	1.28	1.02
			LR	10.18	32.49	93.43	181.50	282.07	380.52	473.55	558.60	638.39	709.11	772.61
7	4	50	Estrella	-0.08	-0.04	0.09	0.29	0.52	0.71	0.84	0.91	0.95	0.97	0.97
3 dominant			BIC	4.13	4.08	3.95	3.71	3.36	2.96	2.54	2.18	1.92	1.77	1.70
			LR	3.99	6.12	12.98	24.81	42.35	62.17	83.08	101.19	114.31	121.59	125.01
7	4	250	Estrella	-0.02	0.02	0.13	0.31	0.51	0.68	0.81	0.90	0.95	0.97	0.98
3 dominant			BIC	3.96	3.93	3.81	3.59	3.30	2.95	2.57	2.21	1.89	1.65	1.54
			LR	4.00	13.23	43.37	96.96	170.34	258.77	352.08	442.47	522.77	583.74	609.17
7	7	50	Estrella	-0.15	-0.08	0.11	0.34	0.57	0.74	0.85	0.92	0.95	0.96	0.97
3 dominant			BIC	4.30	4.23	4.04	3.74	3.37	2.98	2.61	2.29	2.04	1.91	1.82
			LR	7.15	10.49	20.03	34.80	53.70	72.92	91.27	107.46	120.10	126.54	130.77
7	7	250	Estrella*	-0.03	0.04	0.21	0.43	0.65	0.80	0.90	0.95	0.97	0.99	0.99
3 dominant			BIC	4.02	3.95	3.76	3.46	3.07	2.65	2.25	1.91	1.62	1.40	1.23
			LR	7.05	23.14	70.70	147.45	245.35	349.33	448.07	533.58	605.38	661.69	704.80
7	10	50	Estrella*	-0.21	-0.11	0.12	0.41	0.64	0.79	0.88	0.93	0.96	0.98	0.99
3 dominant			BIC	4.47	4.38	4.13	3.76	3.35	2.94	2.58	2.28	2.01	1.80	1.63
			LR	10.22	14.93	27.14	45.59	66.34	86.91	104.72	119.58	133.07	143.61	152.00
7	10	250	Estrella*	-0.04	0.05	0.27	0.51	0.71	0.84	0.92	0.96	0.98	0.99	0.99
3 dominant			BIC	4.07	3.99	3.73	3.37	2.96	2.55	2.18	1.85	1.59	1.37	1.21
			LR	10.20	31.86	94.65	185.97	288.50	390.97	484.21	564.89	631.14	685.73	726.36

(Table continued)

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
10	4	50	Estrella	-0.08	-0.05	0.06	0.20	0.37	0.52	0.65	0.75	0.82	0.88	0.91
			BIC	4.84	4.80	4.69	4.53	4.31	4.06	3.79	3.52	3.28	3.02	2.84
			LR	4.07	5.73	11.31	19.32	30.57	42.83	56.47	69.92	82.10	94.86	104.09
10	4	250	Estrella	-0.02	0.02	0.12	0.26	0.41	0.55	0.67	0.76	0.83	0.87	0.91
			BIC	4.68	4.64	4.53	4.37	4.16	3.92	3.68	3.43	3.19	2.98	2.78
			LR	4.12	13.14	39.87	80.67	133.07	192.25	254.21	315.10	374.89	428.51	479.02
10	7	50	Estrella	-0.15	-0.09	0.10	0.32	0.52	0.67	0.79	0.85	0.90	0.93	0.95
			BIC	5.01	4.95	4.76	4.49	4.16	3.84	3.52	3.27	3.01	2.81	2.66
			LR	7.12	10.04	19.89	33.13	49.65	65.57	81.48	94.38	106.94	117.23	124.47
10	7	250	Estrella	-0.03	0.03	0.20	0.41	0.59	0.74	0.83	0.90	0.93	0.96	0.97
			BIC	4.73	4.67	4.49	4.21	3.88	3.54	3.21	2.90	2.63	2.40	2.24
			LR	7.02	22.93	68.05	137.77	218.72	305.78	388.63	463.78	531.77	589.08	631.18
10	10	50	Estrella	-0.22	-0.12	0.11	0.37	0.60	0.75	0.85	0.91	0.94	0.96	0.96
			BIC	5.18	5.10	4.85	4.52	4.12	3.74	3.38	3.09	2.84	2.69	2.59
			LR	10.17	14.57	26.64	43.20	63.45	82.22	100.16	114.95	127.15	135.04	139.98
10	10	250	Estrella	-0.04	0.05	0.26	0.49	0.68	0.81	0.88	0.93	0.96	0.97	0.98
			BIC	4.79	4.70	4.45	4.11	3.72	3.35	3.01	2.72	2.47	2.29	2.19
			LR	9.95	32.54	93.64	179.20	275.35	368.72	453.09	526.57	589.32	635.18	657.88
10	4	50	Estrella	-0.08	-0.05	0.07	0.22	0.41	0.60	0.76	0.87	0.93	0.96	0.97
j_1 dominant			BIC	4.84	4.80	4.68	4.50	4.24	3.91	3.50	3.08	2.68	2.40	2.27
			LR	4.08	5.77	11.71	20.76	34.09	50.56	70.82	92.03	112.08	125.77	132.52
10	4	250	Estrella	-0.02	0.02	0.13	0.28	0.45	0.62	0.76	0.86	0.93	0.97	0.98
j_1 dominant			BIC	4.68	4.64	4.53	4.35	4.10	3.79	3.43	3.04	2.62	2.23	2.08
			LR	4.20	13.19	41.28	86.44	148.31	225.61	316.23	414.35	517.97	615.32	653.65
10	7	50	Estrella*	-0.15	-0.08	0.13	0.39	0.64	0.81	0.91	0.97	0.99	1.00	1.00
j_1 dominant			BIC	5.01	4.95	4.72	4.38	3.92	3.43	2.92	2.43	1.95	1.55	1.17
			LR	7.12	10.22	21.44	38.50	61.41	86.36	111.86	136.29	159.99	180.07	199.17
10	7	250	Estrella*	-0.03	0.04	0.22	0.46	0.68	0.85	0.94	0.98	0.99	1.00	1.00
j_1 dominant			BIC	4.73	4.67	4.46	4.13	3.68	3.15	2.60	2.08	1.58	1.16	0.82
			LR	7.03	23.53	74.33	157.73	270.20	402.30	539.01	669.88	793.98	898.89	984.03
10	10	50	Estrella*	-0.22	-0.12	0.14	0.46	0.74	0.90	0.97	0.99	1.00	1.00	1.00
j_1 dominant			BIC	5.18	5.09	4.83	4.39	3.78	3.12	2.43	1.85	1.36	1.01	0.83
			LR	10.17	14.72	28.02	49.80	80.20	113.35	148.08	177.03	201.59	219.03	227.91
10	10	250	Estrella*	-0.04	0.05	0.28	0.56	0.80	0.93	0.98	1.00	1.00	1.00	1.00
j_1 dominant			BIC	4.79	4.69	4.42	3.98	3.40	2.72	2.06	1.50	1.06	0.73	0.50
			LR	9.96	33.37	101.25	211.21	357.47	527.31	691.70	831.86	941.69	1023.69	1080.57
10	4	50	Estrella	-0.08	-0.05	0.08	0.25	0.46	0.64	0.79	0.88	0.93	0.95	0.97
5 dominant			BIC	4.84	4.80	4.67	4.47	4.15	3.81	3.40	3.01	2.70	2.47	2.33
			LR	4.07	5.87	12.27	22.64	38.27	55.48	75.72	95.65	110.77	122.52	129.40
10	4	250	Estrella	-0.02	0.02	0.13	0.30	0.48	0.66	0.79	0.88	0.93	0.96	0.97
5 dominant			BIC	4.68	4.64	4.52	4.32	4.04	3.69	3.32	2.96	2.61	2.32	2.14
			LR	4.21	13.29	42.70	93.00	163.17	249.72	343.25	433.65	519.65	593.85	639.56
10	7	50	Estrella*	-0.15	-0.08	0.15	0.42	0.66	0.82	0.91	0.95	0.97	0.98	0.99
5 dominant			BIC	5.01	4.95	4.71	4.34	3.87	3.40	2.96	2.60	2.34	2.14	1.98
			LR	7.12	10.36	22.36	40.51	64.24	87.81	109.62	127.44	140.51	150.80	158.54
10	7	250	Estrella*	-0.03	0.04	0.23	0.48	0.70	0.85	0.93	0.96	0.98	0.99	0.99
5 dominant			BIC	4.73	4.66	4.45	4.09	3.63	3.14	2.70	2.32	2.03	1.82	1.64
			LR	6.90	23.91	77.87	168.56	282.96	405.60	515.73	609.61	681.40	735.92	779.50
10	10	50	Estrella*	-0.22	-0.11	0.19	0.54	0.79	0.90	0.96	0.98	0.99	0.99	0.99
5 dominant			BIC	5.18	5.08	4.76	4.25	3.63	3.09	2.66	2.35	2.12	1.96	1.83
			LR	10.17	15.24	31.37	57.06	87.93	114.65	136.56	151.74	163.45	171.55	178.06
10	10	250	Estrella*	-0.04	0.06	0.32	0.62	0.83	0.93	0.97	0.99	0.99	1.00	1.00
5 dominant			BIC	4.79	4.69	4.38	3.87	3.27	2.71	2.27	1.94	1.71	1.54	1.41
			LR	9.85	34.71	112.70	239.47	389.15	529.43	640.02	720.91	778.26	821.17	853.24

Note: Shaded values indicate statistical significance at the 0.05 level. The asterisk denotes cases in which there was no check for values of Estrella greater than 0.98 and no samples were excluded from the simulation.

The AIC and BIC also improve as we move away from the restricted model. The log likelihood function increases as the signal-to-noise ratio increases, and so does the likelihood-ratio test statistic. In terms of Stein rule estimation the latter implies that smaller weights will be placed on the restricted model, and the Stein rule estimators will converge to the MLE. For sample size N=250 the measures are slightly better, although the differences are not significant. The value of the likelihood-ratio test statistic is bigger and implies faster convergence to the MLE for larger samples, which is what we observe in our study. Changing the direction of the true beta does not alter the goodness of fit measures.

- Squared Error Loss

Table 2.2 shows the relative risk values in terms of squared error loss. The risk of each positive-part Stein estimator is divided by the risk of the MLE, i.e. values smaller than one indicate that shrinkage leads to risk improvement over MLE. The squared error loss measures the distance of the estimator from the true parameter. In other words, this loss measure shows the ability to estimate the parameters of the model. The squared error loss increases as the bias and/or the variance of the estimators increase. The results show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error. The risk of the Stein rule estimators approaches the risk of MLE as the signal-to-noise ratio increases, and the convergence occurs faster for larger samples. The risk improvement is bigger for N=50, which is not surprising, since the performance of the MLE improves as the sample size increases. When N=50, for the parameter space that we consider, with parameter index w increasing from zero to one, the squared error loss of the Stein rule estimator remains smaller than

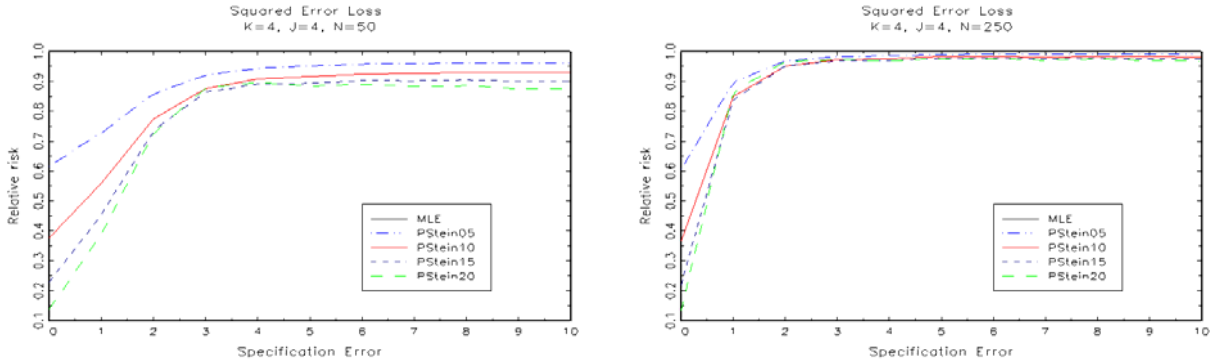
the one of the MLE. When w=1 the parameter vector β reaches its maximum length. When N=50 this corresponds to an Estrella-R^2 of 0.8. When β increases in length still further, the risk of the Stein rule converges to the risk of the MLE. This follows because as $\beta'\beta$ increases, the power of the likelihood ratio test approaches one, so that the Stein rule in equation (2.3) converges to the MLE.

Table 2.2: Squared Error Loss, Four Equally Likely Alternatives

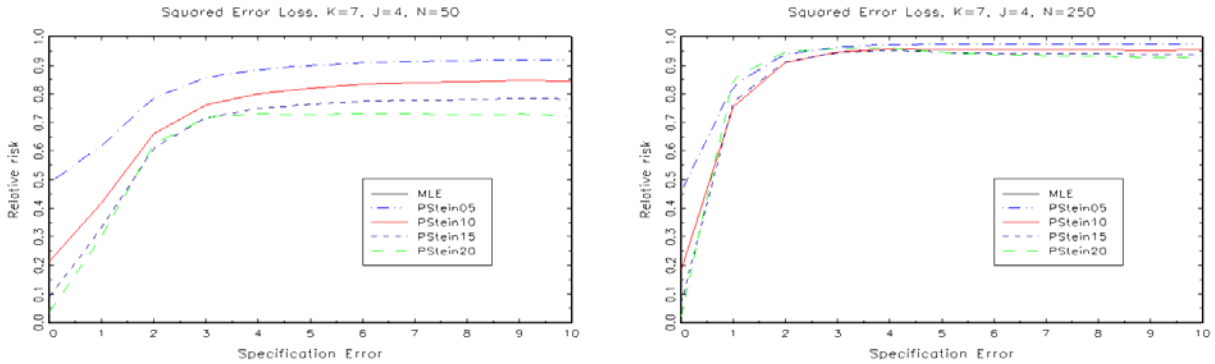
	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	0.12	0.13	0.14	0.14	0.16	0.20	0.22	0.26	0.31	0.38	0.47
	PStein c=0.5	0.62	0.73	0.86	0.92	0.94	0.95	0.96	0.96	0.96	0.96	0.96
	PStein c=1	0.38	0.56	0.77	0.88	0.91	0.92	0.92	0.93	0.93	0.93	0.93
	PStein c=1.5	0.23	0.45	0.73	0.86	0.89	0.89	0.90	0.90	0.90	0.90	0.90
	PStein c=2	0.14	0.39	0.73	0.88	0.90	0.88	0.89	0.88	0.89	0.88	0.88
N=250	Risk MLE	0.02	0.02	0.02	0.03	0.03	0.03	0.04	0.04	0.05	0.06	0.07
	PStein c=0.5	0.61	0.90	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.37	0.85	0.95	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PStein c=1.5	0.22	0.84	0.95	0.97	0.97	0.98	0.98	0.98	0.98	0.97	0.97
	PStein c=2	0.13	0.87	0.97	0.97	0.97	0.98	0.98	0.97	0.97	0.97	0.97
						K=7						
N=50	Risk MLE	0.23	0.23	0.27	0.32	0.39	0.51	0.73	0.96	1.30	1.77	2.18
	PStein c=0.5	0.49	0.62	0.79	0.86	0.89	0.90	0.91	0.91	0.92	0.92	0.92
	PStein c=1	0.22	0.42	0.66	0.76	0.80	0.82	0.83	0.84	0.84	0.85	0.85
	PStein c=1.5	0.09	0.33	0.61	0.72	0.75	0.76	0.77	0.78	0.78	0.78	0.78
	PStein c=2	0.04	0.30	0.62	0.72	0.73	0.73	0.73	0.73	0.73	0.73	0.72
N=250	Risk MLE	0.04	0.04	0.04	0.05	0.06	0.06	0.08	0.09	0.11	0.12	0.15
	PStein c=0.5	0.46	0.83	0.94	0.96	0.97	0.97	0.97	0.97	0.98	0.97	0.97
	PStein c=1	0.19	0.76	0.91	0.95	0.96	0.96	0.95	0.95	0.96	0.95	0.95
	PStein c=1.5	0.07	0.77	0.91	0.94	0.95	0.95	0.94	0.94	0.94	0.94	0.94
	PStein c=2	0.02	0.85	0.95	0.96	0.96	0.95	0.94	0.93	0.93	0.93	0.93
						K=10						
N=50	Risk MLE	0.35	0.36	0.42	0.51	0.61	0.87	1.10	1.64	2.71	3.92	8.21
	PStein c=0.5	0.42	0.56	0.73	0.81	0.84	0.85	0.86	0.87	0.88	0.88	0.90
	PStein c=1	0.14	0.35	0.58	0.69	0.74	0.75	0.75	0.76	0.77	0.78	0.81
	PStein c=1.5	0.04	0.28	0.56	0.66	0.70	0.68	0.68	0.68	0.69	0.69	0.73
	PStein c=2	0.01	0.27	0.61	0.70	0.71	0.66	0.65	0.63	0.62	0.62	0.65
N=250	Risk MLE	0.06	0.06	0.06	0.08	0.09	0.11	0.13	0.15	0.19	0.22	0.27
	PStein c=0.5	0.40	0.80	0.92	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96
	PStein c=1	0.13	0.71	0.89	0.92	0.93	0.93	0.93	0.93	0.93	0.93	0.92
	PStein c=1.5	0.04	0.75	0.90	0.92	0.93	0.92	0.91	0.91	0.91	0.90	0.89
	PStein c=2	0.01	0.88	0.95	0.94	0.94	0.92	0.91	0.90	0.89	0.88	0.87

The risk differences are statistically significant with the exception of Pstein05, the estimator with smallest degree of shrinkage, when Estrella-R^2 exceeds 0.6. Statistical significance at the 0.05 level is indicated by the shaded values in each table. When N=50 more shrinkage offers a larger risk improvement, but there is no dominant Stein Rule estimator over the entire parameter space, because the two estimators with maximum shrinkage cross at w=0.3 which is the weight at which the average value of the LR test statistic becomes statistically significant, and we can reject the null hypothesis of the restricted model. As the sample size increases, Stein rule estimators offer significant gain only over a small range of the parameter space, but the performance of all estimators improves. The likelihood ratio test statistic becomes statistically significant for relatively low levels of Estrella-R^2, as low as 10% when K=4. This is also when the Stein rule estimators begin to converge to the MLE, although the risk of the shrinkage estimators remains slightly lower that the risk of the MLE. In absolute terms, the risk for N=250 is smaller than the risk for N=50 which agrees with asymptotic theory: as the sample size increases, both the bias and the variance of the MLE decrease. The value of the risk increases as K increases which is what we expect, because the squared error loss depends on the number of variables in the model. If we increase the number of variables the relative risk still shows that the Stein estimators dominate the MLE, and overall the risk gain is larger compared to the case of 4 variables. The conventional wisdom tells us that shrinkage works better in cases with more restrictions, which explains why shrinkage estimators perform better in models with more variables. Figure 2.1 shows the plots of the relative risk functions. For clarity of presentation we have plotted only the relative risk of the four positive-rule estimators.

K = 4



K = 7



K = 10

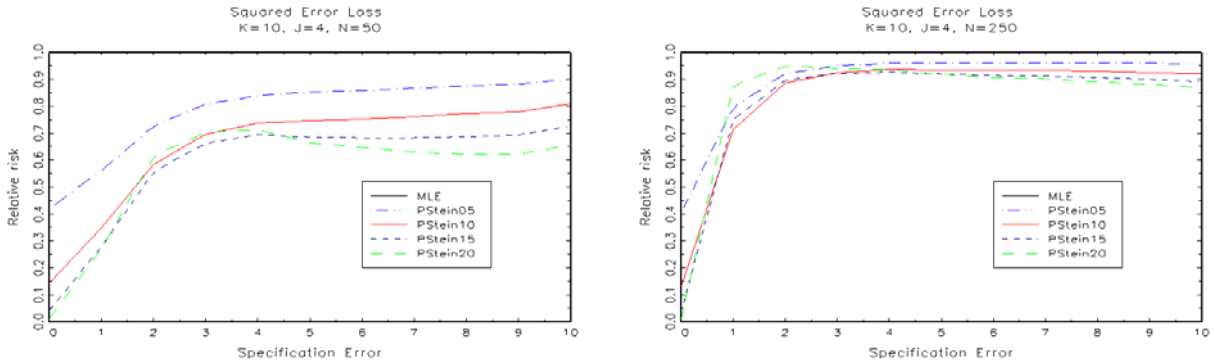


Figure 2.1: Squared Error Loss, Four Equally Likely Alternatives

For *N*=50 the difference between the Stein estimators is more pronounced, showing that bigger shrinkage leads to larger risk improvement, but again there is no dominant Stein rule estimator. For *N*=250 the risk differences improve in terms of statistical significance as the number of variables increases. The model with ten variables

shows larger risk improvement than the model with seven variables, but there are no significant differences between the two cases. Changing the direction of the true parameter vector β does not provide any additional information and will not be discussed separately.

- Weighted Squared Error Loss

Table 2.3 shows the relative risk values in terms of weighted squared error loss. The weighted squared error loss also measures the distance of the estimator from the true parameter and thus the ability to estimate the parameters of the model. The difference from the squared error loss comes from the fact that in this case the loss is weighted by the information matrix of the vector of parameters β. Since the information matrix is the inverse of the covariance matrix, the weighted loss will be inversely related to the variance of the parameters, i.e. more weight will be placed on parameters with smaller variance. In other words, the weighted loss penalizes more for errors in coefficients with smaller variability. Since shrinkage reduces the variability of the estimates, we expect the values of the weighted loss to be higher than the values of the squared loss, and the differences to be more pronounced for smaller degree of specification error. In addition, because variability increases when the coefficients of the model increase in value, the risk increases but the errors will be weighted less compared to cases with smaller values of the coefficients. Therefore, we expect the values of weighted risk to be similar for each degree of hypothesis error. Our results show that the Stein rule estimators dominate the MLE for the entire parameter space, and the risk differences improve in statistical significance as the number of variables increases. The risk improvement of Stein rule estimators over MLE is larger for models with more variables.

Table 2.3: Weighted Error Loss, Four Equally Likely Alternatives

	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	4.31	4.41	4.45	4.45	4.45	4.92	4.84	5.06	5.13	5.47	5.83
	PStein c=0.5	0.62	0.73	0.86	0.92	0.94	0.95	0.96	0.96	0.96	0.96	0.97
	PStein c=1	0.38	0.56	0.78	0.87	0.90	0.91	0.92	0.93	0.93	0.93	0.94
	PStein c=1.5	0.23	0.46	0.74	0.86	0.88	0.88	0.89	0.90	0.90	0.90	0.91
	PStein c=2	0.14	0.40	0.74	0.87	0.88	0.87	0.87	0.87	0.88	0.88	0.88
N=250	Risk MLE	4.05	3.98	4.08	4.00	4.10	4.19	4.22	4.13	4.12	4.16	4.21
	PStein c=0.5	0.61	0.90	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.37	0.86	0.95	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98
	PStein c=1.5	0.22	0.85	0.95	0.97	0.97	0.97	0.98	0.97	0.98	0.98	0.98
	PStein c=2	0.13	0.89	0.97	0.97	0.96	0.97	0.97	0.97	0.97	0.97	0.97
						K=7						
N=50	Risk MLE	8.23	8.06	8.64	8.68	9.04	9.90	11.25	11.98	13.12	14.20	14.40
	PStein c=0.5	0.49	0.63	0.79	0.86	0.88	0.90	0.91	0.92	0.93	0.93	0.93
	PStein c=1	0.21	0.44	0.68	0.76	0.80	0.82	0.84	0.85	0.86	0.87	0.87
	PStein c=1.5	0.09	0.36	0.64	0.72	0.74	0.76	0.78	0.79	0.80	0.81	0.81
	PStein c=2	0.04	0.34	0.68	0.73	0.72	0.72	0.73	0.74	0.75	0.75	0.76
N=250	Risk MLE	7.14	7.27	7.31	7.26	7.28	7.32	7.37	7.57	7.57	7.62	7.67
	PStein c=0.5	0.46	0.83	0.94	0.96	0.97	0.97	0.98	0.98	0.98	0.98	0.98
	PStein c=1	0.19	0.76	0.90	0.94	0.95	0.95	0.96	0.96	0.96	0.97	0.97
	PStein c=1.5	0.07	0.78	0.90	0.93	0.94	0.94	0.94	0.95	0.95	0.95	0.95
	PStein c=2	0.02	0.87	0.93	0.93	0.93	0.93	0.93	0.93	0.94	0.94	0.94
						K=10						
N=50	Risk MLE	11.8	12.1	12.8	13.6	13.9	15.9	16.7	19.3	23.4	26.5	37.5
	PStein c=0.5	0.42	0.56	0.72	0.80	0.84	0.86	0.87	0.89	0.89	0.90	0.91
	PStein c=1	0.14	0.35	0.57	0.67	0.72	0.75	0.77	0.79	0.80	0.81	0.83
	PStein c=1.5	0.04	0.29	0.54	0.61	0.65	0.67	0.68	0.70	0.72	0.73	0.75
	PStein c=2	0.01	0.28	0.58	0.61	0.61	0.61	0.62	0.63	0.65	0.66	0.68
N=250	Risk MLE	10.4	10.4	10.1	10.6	10.5	11.0	10.7	11.0	11.2	11.1	11.4
	PStein c=0.5	0.40	0.80	0.92	0.95	0.96	0.96	0.97	0.97	0.97	0.97	0.97
	PStein c=1	0.13	0.72	0.88	0.91	0.93	0.94	0.94	0.95	0.95	0.95	0.95
	PStein c=1.5	0.04	0.75	0.87	0.90	0.91	0.91	0.92	0.92	0.93	0.93	0.93
	PStein c=2	0.01	0.88	0.91	0.90	0.90	0.90	0.90	0.91	0.91	0.91	0.91

The absolute risk is smaller in large samples, and it increases as *K* increases for each sample size. The risk increases as the coefficients increase in values, but as expected, the differences are relatively small, especially for *N*=250. The differences are larger for models with 50 observations, especially when the number of variables increases, because adding more variables make precise estimation more difficult in small

samples. In addition, for each model, the values of the weighted risk for N=50 are not very different from the values for N=250 unlike the case of squared error loss, but the difference increases as K increases. The risk functions are presented in Figure 2.2.

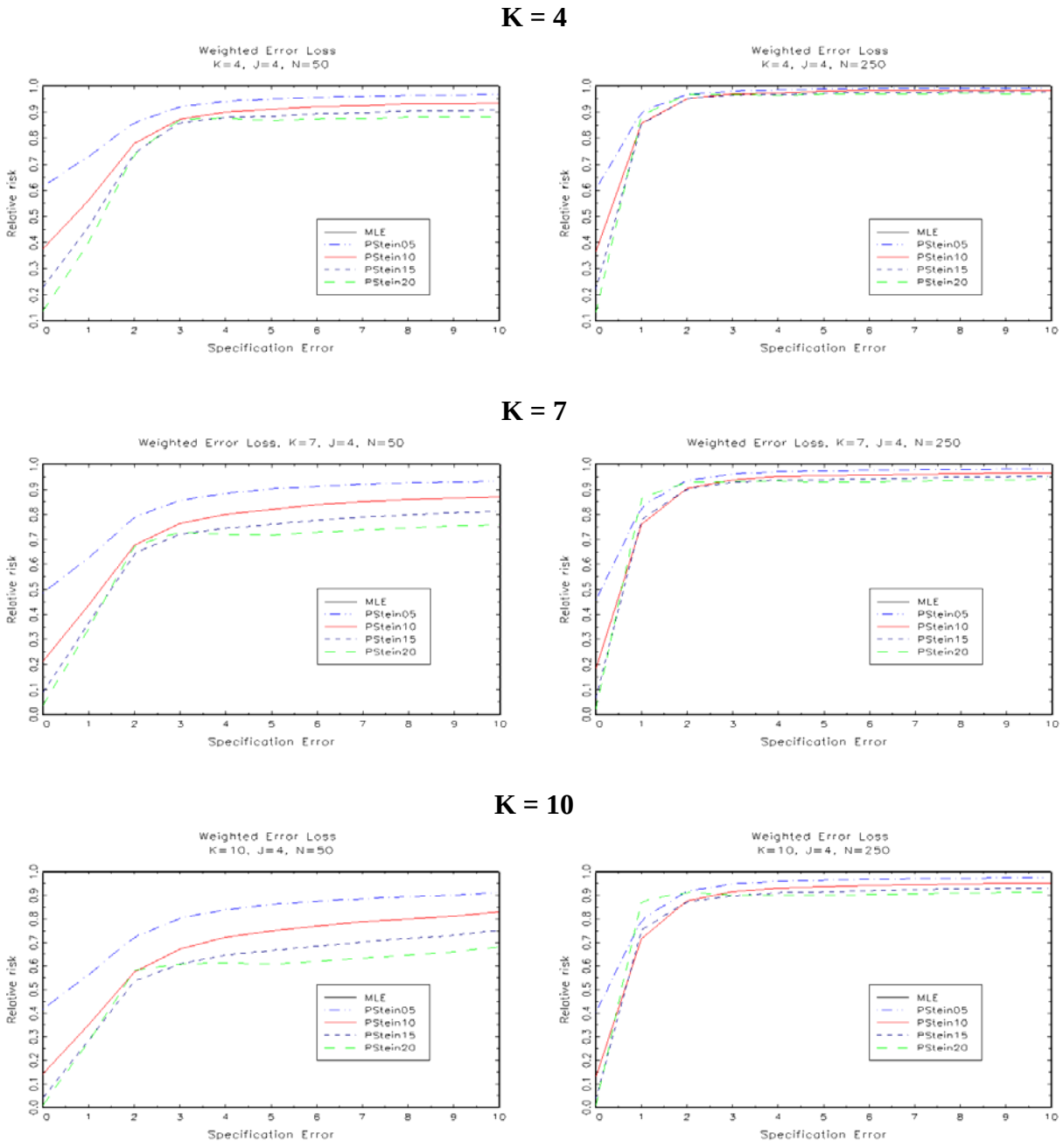


Figure 2.2: Weighted Squared Error Loss, Four Equally Likely Alternatives

More shrinkage leads to larger risk improvement in small samples, because it reduces the absolute magnitude of the parameter estimates, as compared to the unrestricted estimator, which reduces their variability. In large samples the variance of all estimators decreases which diminished the benefits of shrinking the coefficients. Also here, there is no dominant Stein rule estimator. As we explained earlier, there is a difference is the actual risk values compared to the squared loss – the weighted loss is larger than the squared loss, because the weighted loss takes into account the variance of the coefficients and pays more attention to errors in precisely estimated coefficients. Again, changing the true beta does not change the results.

- Marginal Effects

The results for the relative risk of the marginal effects are similar to the ones obtained for the squared and weighted error loss, and are not reported, but can be provided upon request[8]. The Stein rule estimators dominate the MLE over the entire parameter space, with significant risk improvement in small samples and for small degrees of specification error. There is no dominant Stein rule estimator, although in most cases more shrinkage means larger risk improvement. The numerical values of the risk here are very small, and, as expected, they are smaller for N=250 than for N=50. Charging the true beta does not affect the results.

- In-Sample Prediction Loss

Next we evaluate the models in terms of predictive ability. The true probabilities are computed as $P_{ij} = \frac{\exp(z_{ij}'\beta)}{\sum_{j=1}^{J}\exp(z_{ij}'\beta)}$, and the predicted probabilities are computed as

[8] If additional results are needed, please contact the author (vtabak1@lsu.edu) or the LSU Department of Economics.

$\hat{P}_{ij} = \frac{\exp(z'_{ij}\hat{\beta})}{\sum_{j=1}^{J}\exp(z'_{ij}\hat{\beta})}$, where $\hat{\beta}$ takes the values obtained by each of the nine estimators. The loss function is shown in equation (2.13). It measures the squared distance between the true and predicted probability. The values of the prediction risk are very small, because the difference in probabilities $P_{ij} - \hat{P}_{ij}$ is a very small number, which is then squared and summed over the number of alternatives *J*. The obtained value is summed over *i* but then divided by the number of observations *N*. We report the average values over all Monte Carlo samples. The risk of the MLE for *N*=50 ranges from 0.02 for *K*=4 to 0.06 for *K*=10. As we increase the sample size, these values get smaller. We expect good performance of the MLE for prediction in sample, because it chooses the parameters that maximize the probability of obtaining the sample of observations, and we use the same sample to determine the probabilities of choosing an alternative. The results for relative risk show that shrinkage estimators still dominate the MLE for small degrees of specification error, but convergence to the MLE occurs fast, especially in large samples, usually for values of Estrella-R^2 lower than 10%. The risk improvement over MLE is greater as the number of variables increases. Another way of evaluating forecasts is looking at the percentage of correct hits for each estimator. The value of this measure is called the "hit rate" and is calculated as $p_{11} + ... + p_{jj} + ... + p_{JJ}$, which is the sum of the percent correctly predicted outcomes for each alternative. The hit rate increases as we increase the signal-to-noise ratio and reaches a value of .7 or higher for all models when the Estrella-R^2 exceeds .8. There are no differences in the hit rate obtained by the different estimators. Changing the true beta does not provide any additional information and does not change the results.

- Out-Of-Sample Prediction Loss

The results for prediction out-of-sample are presented in Table 2.4.

Table 2.4: MSE of Prediction Out of Sample, Four Equally Likely Alternatives

	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	2.08	2.15	2.17	2.15	2.11	2.24	2.15	2.17	2.16	2.23	2.36
	PStein c=0.5	0.62	0.74	0.88	0.95	0.97	0.98	0.98	0.98	0.99	0.99	0.99
	PStein c=1	0.38	0.57	0.81	0.92	0.95	0.96	0.97	0.97	0.98	0.98	0.98
	PStein c=1.5	0.23	0.48	0.79	0.93	0.96	0.96	0.97	0.97	0.97	0.97	0.97
	PStein c=2	0.14	0.42	0.79	0.96	0.98	0.97	0.97	0.96	0.97	0.97	0.97
N=250	Risk MLE	0.45	0.44	0.45	0.45	0.47	0.48	0.49	0.48	0.48	0.49	0.51
	PStein c=0.5	0.61	0.90	0.97	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	PStein c=1	0.37	0.86	0.96	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1.5	0.22	0.85	0.96	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=2	0.13	0.88	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99
						K=7						
N=50	Risk MLE	3.90	3.82	4.16	4.30	4.66	5.06	5.72	6.00	6.39	6.84	7.18
	PStein c=0.5	0.50	0.64	0.82	0.89	0.92	0.94	0.96	0.97	0.97	0.97	0.98
	PStein c=1	0.22	0.44	0.70	0.81	0.86	0.90	0.92	0.93	0.94	0.95	0.96
	PStein c=1.5	0.09	0.35	0.66	0.77	0.82	0.86	0.88	0.90	0.92	0.93	0.94
	PStein c=2	0.04	0.32	0.67	0.77	0.79	0.83	0.86	0.88	0.89	0.91	0.92
N=250	Risk MLE	0.70	0.72	0.74	0.75	0.74	0.74	0.72	0.74	0.74	0.73	0.74
	PStein c=0.5	0.46	0.83	0.95	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.19	0.76	0.93	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99
	PStein c=1.5	0.07	0.77	0.94	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PStein c=2	0.02	0.85	0.97	0.99	0.99	0.98	0.98	0.98	0.98	0.98	0.98
						K=10						
N=50	Risk MLE	5.61	5.65	5.85	5.97	5.77	5.75	5.65	5.72	5.82	5.88	5.84
	PStein c=0.5	0.44	0.60	0.79	0.88	0.93	0.94	0.96	0.96	0.97	0.97	0.97
	PStein c=1	0.15	0.39	0.67	0.82	0.88	0.91	0.92	0.93	0.94	0.95	0.95
	PStein c=1.5	0.04	0.32	0.66	0.81	0.87	0.89	0.91	0.91	0.92	0.93	0.93
	PStein c=2	0.01	0.32	0.75	0.88	0.92	0.91	0.91	0.91	0.91	0.91	0.91
N=250	Risk MLE	1.12	1.11	1.11	1.19	1.19	1.23	1.19	1.22	1.23	1.22	1.23
	PStein c=0.5	0.41	0.80	0.93	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.13	0.71	0.89	0.95	0.96	0.97	0.98	0.98	0.98	0.98	0.99
	PStein c=1.5	0.04	0.72	0.89	0.94	0.96	0.96	0.97	0.97	0.98	0.98	0.98
	PStein c=2	0.01	0.83	0.92	0.95	0.96	0.96	0.97	0.97	0.97	0.97	0.98

For each Monte Carlo experiment we generate a sample of 100 observations which are not used in the estimation of the model parameters. The data generating process is

the same as the one used in the original sample. The true probabilities are computed as $P_{ij} = \frac{\exp(z'_{Oij}\beta)}{\sum_{j=1}^{J}\exp(z'_{Oij}\beta)}$, and the predicted probabilities are computed as $\hat{P}_{ij} = \frac{\exp(z'_{Oij}\hat{\beta})}{\sum_{j=1}^{J}\exp(z'_{Oij}\hat{\beta})}$, where $\hat{\beta}$ takes the values obtained by each of the nine estimators using the original sample, and Z_O is the matrix of observations generated for out-of-sample prediction. We use two different loss functions to evaluate out-of-sample performance, MSE_P and MSE_Y, shown in equations (2.13) and (2.14). The results in the table are obtained using MSE_P. The values of risk are very small, as explained in the previous section, therefore the risk of the MLE is multiplied by 100 in the output tables to be able to make comparison between models. The results for relative risk show that shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. Stein rule estimators show larger risk improvement as the number of variables increases. For *N*=50, the shrinkage estimators converge to the MLE when the Estrella-R^2 is close to .2 for the model with 4 variables, and higher than .3 for the models with more variables. For *N*=250, convergence occurs for values of Estrella-R^2 equal to .11, .17, and .22 for *K*=4, 7 and 10 respectively. If we use MSE_Y, we replace the true probability with the actual choice of an alternative, where $y_{ij} = 1$ for the alternative with the highest utility, and zero otherwise. In this case our conclusions do not change, but the differences between the relative risks are very small. The value of risk is higher under this measure, but comparison cannot be made, because in this case we subtract a probability from a dummy variable. The hit rate shows very good predictive ability out-of-sample and increases as we increase the signal-to-noise ratio. The values of the hit rate are higher in-sample than

out-of-sample, but the differences are small. This is a result we expect because the estimated coefficients $\hat{\beta}$ are based on the original sample, while the out-of-sample explanatory variables and choice variables are not used in the estimation process. Overall, the results for prediction do not differ from the results for estimation and we should use the same considerations when choosing an estimator, regardless the main purpose of our research. The plots of the risk functions are similar to the ones for squared and weighted error loss and are not reported.

2.4.2 Four Alternatives With One Dominant

The model with one dominant alternative represents cases in which one of the alternatives is chosen by the majority of individuals. In marketing context such a model describes a situation in which one of the brands is the market leader, and all other brands have only a small market share. This is an interesting example, because such a situation can be encountered often in practical applications. An alternative may be preferred over the other alternatives if one or more of the explanatory variables have values significantly different from the values of other alternatives. For example, a brand may be preferred because it offers the lowest price, or because it invests more heavily in advertising compared to the competing brands. Following this logic, a dominant alternative was created by changing the mean of the explanatory variables associated with this alternative. It was achieved in the Monte Carlo experiment by adding a constant to the explanatory variables of the dominant alternative. In most cases it was sufficient to add 0.1 to the standard normal variables to generate an alternative which was selected by 80% of the individuals when the values of the true β equal to one. For smaller degrees of hypothesis error the share of the dominant alternative is smaller, and all alternatives are

equally likely when specification error is zero and the true β is determined only by the restricted model.

- Goodness of Fit

Table 2.1 shows the goodness of fit measures. The measures improve compared to the model with four equally likely alternatives, and the improvement is larger as we increase the signal-to-noise ratio, which corresponds to larger share of the dominant alternative. This is because as the probability of choosing an alternative increases, the value of the log-likelihood increases as well. As the true β increases in value, the likelihood ratio test statistics get bigger, which shows that the hypothesis of equally likely alternatives is rejected a larger percent of the time. As we increase the number of variables, the values of the Estrella-R^2 get close to one. The chances of a probability being equal to one or to zero increase, which makes the model a perfect predictor and may cause the estimation to break down. To avoid this problem we excluded from the Monte Carlo experiment samples with values of the Estrella-R^2 greater than 0.98, which in the models with seven and ten variables lead to a number of exclude samples exceeding the acceptable level of 10%. The exact number of excluded samples for each model is shown in Table 2.5 at the end of Section 2.4.

- Squared Error Loss

The results for relative risk are very similar to the model with equally likely alternatives and show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error. Shrinkage leads to risk improvement also in this case due to the following reasons. For small degrees of hypothesis error the parameters of the model are close to zero and all alternatives have

similar shares. In particular, each alternative get chosen 25% of the time when the true β is a vector of zeros. In these cases we place more weight on the restricted estimator which significantly reduces estimation risk. As the model coefficients increase in value the differences between the shares of each alternative increase as well. In these cases the value of the likelihood ratio test statistic is large and we place more weight on the unrestricted estimator, therefore we are not imposing the restriction of equally likely alternatives. Generally, when the share of the dominant alternative exceeds 30%, the likelihood ratio test statistic is statistically significant and we can reject the hypothesis of a null vector of coefficients. As in the case with equally likely alternatives, the results show that the risk improvement is larger for N=50 and for models with more explanatory variables. The risk differences for N=50 are statistically significant with the exception of Pstein05, the estimator with smallest degree of shrinkage, when the Estrella-R^2 exceeds 0.6 for the model with K=4, and 0.8 for the model with K=10. When the number of variables increases, the risk differences are statistically significant also for N=250, again with the exception of Pstein05 for larger degrees of specification error. Generally more shrinkage offers a larger risk improvement, but there is no dominant Stein Rule estimator over the entire parameter space. Convergence to the MLE occurs for values of Estrella-R^2 as low as 0.10 in the large sample, and as high as .34 in the small sample, which in most cases coincides with the value of the likelihood ratio test statistic becoming statistically significant. The value of the risk is higher compared to the model with equally likely alternatives, although the differences are small. We also observe an increase in the estimator bias[9] which is a reason for the risk increase. The bias is

[9] We only estimate the bias of the first coefficient.

calculated as the difference between the true parameter β and the average value of the estimated coefficients over all Monte Carlo samples. The bias is not reported, but the results can be provided upon request. As we expect, the values for the risk are smaller in large samples and for models with smaller number of explanatory variables.

- Weighted Squared Error Loss

Our results for weighted error loss are similar to the case with four equally likely alternatives and show that the Stein rule estimators dominate the MLE for the entire parameter space, and the risk differences improve in statistical significance as the number of variables increases. The risk improvement of Stein rule estimators over MLE is larger in small samples and for models with more variables. More shrinkage leads to larger risk improvement, but there is no dominant Stein rule estimator. Like in the case of squared error loss, the Stein rule estimators converge to the MLE for small values of Estrella-R^2, on average close to .20 for N=250 and to .3 for N=50. The values of the weighted risk are similar to the values when all alternatives are equally likely, and decrease as we increase the number of observations. The risk differences between the two samples sizes get larger as we increase the number of variables, because as the model complexity increases 50 observations are not sufficient for precise estimation and we observe substantial risk increase for higher degrees of specification error. Again, for the model with four variables the risk differences between weights are small, and they increase for K=7 and K=10.

- Marginal Effects

The results for the relative risk of the marginal effects are similar to the ones obtained for the squared and weighted error loss, and similar to the results for the case

with equally likely alternatives. Shrinkage decreases the variability of the estimates and decreases the risk, and the gains are larger in small samples, because as we increase the number of observations the bias and variance decrease and the performance of the maximum likelihood estimator improves. The Stein rule estimators dominate the MLE over the entire parameter space, but again there is no dominant Stein rule estimator, although in most cases more shrinkage means larger risk improvement.

- In-Sample Prediction Loss

The values of the prediction risk are very small and very similar for each degree of specification error. The MLE performs very well while predicting the probabilities in sample, and the risk gain over MLE is significant only for values of β close to zero, when more weight is placed on the restricted estimator. The hit rate has the same values for all estimators for values of Estrella-R^2 higher than 10%. In addition, the hit rate is higher compared to the model with equally likely alternatives. This result is logical because it is easier to predict outcomes when there is only one dominant alternative. The hit rate increases as we increase the number of variables, because we add more information relevant to the choice of an alternative. Looking at the actual shares and the predicted shares for each alternative shows very small differences, which also confirms that the estimators perform very well as predictors in-sample. Generally, both the MLE and the Stein rule estimators over-predict the dominant alternative for all degrees of specification error, but the differences between actual and predicted shares get smaller as we increase the signal-to-noise ratio.

- Out-Of-Sample Prediction Loss

The results for relative risk of out-of-sample prediction show that shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. Stein rule estimators show bigger risk improvement as the number of variables increases. For N=50, the shrinkage estimators converge to the MLE when Estrella-R^2 is close to .2 for the model with 4 variables, and higher than .3 for the models with more variables. For N=250, convergence occurs for values of Estrella-R^2 between .1 and .2 depending on the number of variables in the model. More shrinkage results in larger risk improvement but the risk of the maximum shrinkage estimator Pstein20 exceeds the risk of the other shrinkage estimators upon convergence to MLE, and therefore, there is no dominant estimator over the entire parameter space. The hit rate shows very good predictive ability out-of-sample and exceeds 80% at the highest signal-to-noise ratio. Comparable to the in-sample hit rate, the percent correctly predicted outcomes is higher than in the case of equally likely alternatives. The values of the hit rate for all estimators are the same. The actual and predicted shares out-of-sample show that both the MLE and the Stein rule estimators tend to over-predict the dominant alternative, a result similar to the case of prediction in-sample. Overall, all estimators predict very well out-of-sample, but the Stein estimators outperform the MLE for small to moderate degrees of hypothesis error.

2.4.3 Four Alternatives With Two Dominant

The model with two dominant alternatives represents cases in which two of the alternatives are chosen by the majority of individuals, and the two alternatives have similar shares. In marketing context such a model describes a situation in which we have

two major brands that compete with each other, and the remaining brands have only a small market share. One of the examples in the fourth chapter discusses the choice of four brands of saltine crackers, where two of the brands (Nabisco and a Private label) together get selected about 86% of the time, and the other two brands (Sunshine and Keebler) share the remaining 14%. We use the same reasoning about the choice of an alternative as in the previous case, and assume that the explanatory variables associated with the dominant alternatives will have different mean from the rest of the variables. Creating two dominant alternatives was achieved in the Monte Carlo experiment by adding a constant equal to 0.1 to the standard normal explanatory variables of the dominant alternatives. The common share of the two dominant alternatives is about 90% when the values of the true β equal to one. Like in the previous case, for smaller degrees of hypothesis error the shares of the dominant alternatives are smaller, and all alternatives are equally likely when specification error is zero and the true β is determined only by the restricted model.

- Goodness of Fit

The goodness of fit measures are comparable to the model with one dominant alternative, although the values are a little lower because the chances of a probability being equal to one is much smaller in this case. The number of excluded samples is generally smaller compared to the model with one dominant alternative, but the weights for which the Monte Carlo experiment breaks down overlap in all cases except for the model with N=250 and K=10.

- Squared Error Loss

The results for relative risk are very similar to the models with one dominant or equally likely alternatives and show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error, for the same reasons discussed in the case of one dominant alternative. At low degree of specification error we place more weight on the restricted estimator, but the actual shares are very similar, and identical when $\beta = 0$. As the differences in shares for each alternative increase we weigh heavily the unrestricted model and do not impose the restriction of equally likely alternatives. The risk differences for *N*=50 are statistically significant with the exception of Pstein05, the estimator with smallest degree of shrinkage, when Estrella-R^2 exceeds 0.6 for the model with *K*=4, and 0.8 for the model with *K*=10, and Pstein10, the base model, when *K*=4 and Estrella-R^2 equals .84. When the number of variables equals 7 or 10, the risk differences are statistically significant also for *N*=250, again with the exception of Pstein05 for *K*=7 and larger degrees of specification error. Generally more shrinkage offers a larger risk improvement, but again the risk of the maximum shrinkage estimator exceeds the risk of one or more of the remaining shrinkage estimators before the risk function levels out (or converges to the MLE for *N*=250) and therefore no Stein Rule estimator has lower risk over the entire parameter space. Convergence to the MLE occurs for values of Estrella-R^2 identical to the values of the model with one dominant alternative. The values of the risk are almost identical to the model with one dominant alternative as well. As we expect, the values of risk are smaller in large samples and for models with less explanatory variables.

- Weighted Squared Error Loss

As in the previous two cases Stein rule estimators dominate the MLE for the entire parameter space, and the risk improvement is larger in small samples and for models with more variables. More shrinkage leads to larger risk improvement, but the risk of Pstein20 is higher than the risk of Pstein15 for K=4, and also higher than the risk of Pstein10 for K=7 and 10 when the risk functions of all estimators level out, which occurs for values of Estrella-R^2 lower than .10 when N=50 and lower than .05 when N=250. After that it decreases again and remains below the risk of the other estimators as the hypothesis error increases. The values of the weighted risk are very similar to the values when one alternative is dominant, and decrease as we increase the number of observations. Also here the risk differences between the two samples sizes get bigger as we increase the number of variables.

- Marginal Effects

The results for the relative risk of the marginal effects are similar to the ones obtained for the cases with one dominant or four equally likely alternatives. The Stein rule estimators dominate the MLE over the entire parameter space, but again there is no dominant Stein rule estimator because of the behavior of the maximum-shrinkage estimator Pstein20.

- In-Sample Prediction Loss

As in the previous two cases, the values of the prediction risk are very small and very similar for each degree of specification error. The MLE performs very well while predicting the probabilities in sample, and the risk gain over MLE is significant only for values of β close to zero, when more weight is placed on the restricted estimator.

Convergence to the MLE occurs fast, for values of Estrella-R^2 less than .10 if N=50, and less than 0.03 in N=250. The hit rate equals .9 when K=10 and $\beta = 1$, and has the same values for all estimators for values of Estrella-R^2 higher than .10. Compared to the previous two cases, the hit rate is larger than in the models with equally likely alternatives and smaller than in the model where one alternative is dominant. The actual shares and the predicted shares for each alternative also confirm that all estimators perform very well as predictors in-sample.

- Out-Of-Sample Prediction Loss

The results for relative risk of out-of-sample prediction confirm out findings in the previous two cases: shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. For N=50, the shrinkage estimators converge to the MLE when Estrella-R^2 is close to .2 for the model with 4 variables, and higher than .3 for the models with more variables. For N=250, convergence occurs for values of Estrella-R^2 equal to .11, .17 and .22 for K=4, 7 and 10 respectively. The hit rate shows very good predictive ability out-of-sample and exceeds .8 for some models at the highest signal-to-noise ratio. The values of the hit rate for all estimators are the same and there are no significant differences as we increase the sample size or the number of variables in the model. Overall, also in this case, all estimators predict very well out-of-sample, but the Stein estimators outperform the MLE for small to moderate degrees of hypothesis error.

2.4.4 Seven or Ten Equally Likely Alternatives

Our next objective is to explore the performance of the estimators when the number of alternatives increases. We choose models with 7 or 10 alternatives, which is a

reasonable assumption because in reality often many brands compete for the same target market. We want to confirm our findings about relative risk and to determine whether more variables or more alternatives favor shrinkage estimation.

- Goodness of Fit

The goodness of fit measures generally increase when the number of alternatives increases. This also increases the number of samples excluded from the experiments, and the highest number of excluded samples is for the models with J=10 and K=10. The values of the likelihood-ratio test statistic generally are higher compared to the relevant model with smaller number of alternatives, which implies that the assumption of all alternatives being equally likely is more likely to be incorrect in models with large number of alternatives. In terms of Stein rule estimation the former implies that smaller weights will be placed on the restricted model.

- Squared Error Loss

The results show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error. The risk differences, however, are larger for models with smaller number of alternatives, when the sample size is small. As the sample size increases, we do not notice differences in terms of relative risk when we change the number of variables and/or the number of alternatives. In this case the maximum shrinkage estimator Pstein20 converges faster to the MLE for small degree of specification error and corresponding values of Estrella-R^2 below .10, and then its risk becomes lower than the risk of the other shrinkage estimators. This behavior is consistent for all models and implies that there is no dominant Stein rule estimator over the entire parameter space. For the model with seven alternatives, the risk differences are

statistically significant if N=50 with the exception of Pstein05 when K=4. For N=250, the risk differences improve in statistical significance as K increases. We observe the same pattern for the model with ten alternatives. In absolute terms there is a risk improvement in all models as we increase the number of alternatives.

- Weighted Squared Error Loss

The values of the weighted loss are higher than the values of the squared loss because in this case the errors in coefficients with lower variability are more heavily weighted, and shrinkage reduces the variability of the estimates. The values of the weighted loss for different levels of hypothesis error are very similar, and so are the values in small and large samples, which is what we expect for the information weighted error loss. Also in this case the value of risk decreases as J increases, but the differences are relatively small. In terms of relative risk, our results show that the Stein rule estimators dominate the MLE for the entire parameter space, and the risk differences improve in statistical significance as the number of variables increases. The risk improvement of Stein rule estimators over MLE is larger for models with more variables, but does not increase as we increase the number of alternatives. Also here, there is no dominant Stein rule estimator because of the increase in risk of the maximum shrinkage estimator when the risk of all estimators approaches the risk of MLE.

- Marginal Effects

The results for the relative risk of the marginal effects are similar to the ones obtained for the squared and weighted error loss, and for the models with four alternatives. The Stein rule estimators offer risk improvement over MLE for all degrees

of specification error. The numerical values of the risk here are very small as well, and decrease with the increase in number of observations.

- In-Sample Prediction Loss

The results for relative risk show that shrinkage estimators still dominate the MLE for small degrees of specification error, but convergence to the MLE occurs fast, especially in large samples, usually for values of Estrella-R^2 lower than .5. These results are consistent with our finding for the model with 4 alternatives and with the fact that the maximum likelihood estimator is a good predictor in-sample. The hit rate, which gives the sum of the percent correctly predicted outcomes, is lower compared to the model with four alternatives. The possibility to make a mistake in predicting the choice of an alternative should increase as we add to the model more outcomes that need to be predicted. As in the case with four alternatives, there are no differences in the hit rate obtained by the different estimators.

- Out-Of-Sample Prediction Loss

The results for relative risk show that shrinkage improves prediction, and for N=50 the shrinkage estimators converge to the MLE for larger degree of specification error compared to the model with four alternatives. This corresponds to values of Estrella-R^2 ranging from .36 for K=4 to .55 for K=10 with seven alternatives, and from .37 to .68 in the corresponding cases for ten alternatives. Overall the risk values are lower than when J=4. Also here Stein rule estimators show larger risk improvement as the number of variables increases. Overall more shrinkage leads to larger risk improvement but no estimator dominates the others over the entire parameter space. The hit rate is similar to the hit rate in-sample and the values are lower compared to models

with less alternatives. In this case, the results show that an increase in the number of alternatives favors shrinkage in prediction out-of-sample but does not show risk improvement for estimation compared to models with smaller number of alternatives.

2.4.5 Seven Or Ten Alternatives With One Dominant

As in the case of four alternatives a dominant alternative was created by changing the mean of the standard normal explanatory variables associated with it. It was achieved by adding a constant to these variables, and the same constant was used in all models. For smaller degrees of hypothesis error the share of the dominant alternative is smaller, and all alternatives are equally likely when specification error is zero and the true β is determined only by the restricted model. When β is determined only by the unrestricted model, the share of the dominant alternative exceeds 70% in most models.

- Goodness of Fit

The values of Estrella-R^2 increase compared to the models where all alternatives are equally likely, and approaches one as β increases in value. The chances of a probability being equal to one or to zero increase even more when the number of alternatives is high. This is a problem and may cause the estimation to break down. The number of excluded samples significantly increases and is the highest for models where K=10 and J=10. In several cases, the programs had to run without a check for the values of Estrella-R^2 because of the high computation time.

- Squared Error Loss

The results for relative risk are very similar to the model with equally likely alternatives and show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error. However, the risk

improvement is larger for models with smaller number of alternatives, although there is no significant difference in terms of relative risk between the models. As we increase the number of alternatives, the absolute risk decreases, hence the performance of all estimators improves. As in all previously discussed cases, the results show that the risk improvement is larger for N=50 and for models with more explanatory variables, and also here no Stein rule estimator has the lowest risk over the entire parameter space.

- Weighted Squared Error Loss

Our results for relative risk show that the Stein rule estimators dominate the MLE for the entire parameter space. The risk is lower for models with large number of alternatives as long as the values of Estrella-R^2 stay below .95. As the models increase and the values of Estrella-R^2 approach one, we observe a severe risk increase for large degrees of specification error, especially when w=1. Such values can be observed for the model with ten alternatives when K=7 and K=10. Since it happens in cases where the number of excluded samples exceeds the allowed 10% we will not consider these risk values in our analysis.

- Marginal Effects

The results for the relative risk of the marginal effects do not provide any additional information as a result of the increase in the number of alternatives. Shrinkage estimators have lower risk than the MLE, and the gains are larger in small samples, because as we increase the number of observations the bias and variance decrease and the performance of the maximum likelihood estimator improves.

- In-Sample Prediction Loss

The values of the prediction risk are very small and very similar for each degree of specification error. The MLE performs very well while predicting the probabilities in sample, and the risk gain over MLE is significant only for values of β close to zero, when more weight is placed on the restricted estimator. The hit rate is higher compared to the model with equally likely alternatives. However, the actual and predicted shares in sample show that both the Stein estimators and the MLE over-predict the dominant alternative and the gap is larger compared to the model with four alternatives.

- Out-Of-Sample Prediction Loss

In terms of out-of-sample prediction in this case, comparable to the other measures, we should not consider the results of the experiment for values of Estrella-R^2 greater than .95 because of the large number of excluded samples. The results for relative risk show that shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. The risk decreases in value as the number of alternatives increases. The hit rate shows improved predictive ability compared to the model with equally likely alternatives. Also in this case, both the MLE and the shrinkage estimators over-predict the dominant alternative.

2.4.6 Seven Or Ten Alternatives With Half Of Them Dominant

The model considers three dominant alternatives when J=7 and five dominant alternatives when J=10. Usually for large degrees of specification error the share of the dominant alternatives exceeds 90% and 10% or less is distributed among the remaining alternatives. This increases the risk of getting a value of zero for the probability of an alternative, which would affect the likelihood function and, hence, the risk of all

estimators. Creating the dominant alternatives was achieved in the same way as in all previously described models.

- Goodness of Fit

The goodness of fit measures are comparable to the models with one dominant alternative and approach one if the programs are executed in their original form. The number of excluded samples is comparable to the model with one dominant alternative, although in some instances it is smaller.

- Squared Error Loss

The results for relative risk confirm our conclusions in others models that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error. The risk differences are statistically significant in most cases, and the exceptions are usually for the estimator with minimum shrinkage Pstein05. Convergence occurs for the same values of Estrella-R^2 as in the model with equally likely alternatives and with one dominant alternative.

- Weighted Squared Error Loss

The risk jump for weighted error loss is much smaller here than in the model with one dominant alternative. As in the previous two cases Stein-rule estimators dominate the MLE for the entire parameter space, and the risk improvement is bigger in small samples and for models with more variables. The values of the weighted risk get smaller as J increases, although in most cases the risk differences are very similar between models. Also here the risk differences between the two samples sizes get larger as we increase the number of variables.

- Marginal Effects

The results for the relative risk of the marginal effects are similar to the ones obtained for the cases with one dominant or four equally likely alternatives, and for smaller number of alternatives. The Stein rule estimators dominate the MLE over the entire parameter space, but again there is no dominant Stein rule estimator.

- In-Sample Prediction Loss

As in the previous two cases, the risk gain for the shrinkage estimators is substantial only when the values of β are close to zero. Convergence to the MLE occurs fast, for average values of Estrella-R^2 close to .1 in N=50 and less than .05 if N=250. The value of the hit rate is lower than in the case of one dominant alternative, but is still above 60% for most of the models.

- Out-Of-Sample Prediction Loss

The results for prediction out-of-sample in terms of relative risk confirm our findings in the previous two cases: shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. For N=250, the shrinkage estimators converge to the MLE when Estrella-R^2 is close to .2 on average. For N=50 convergence occurs towards the mid range of parameter space. The hit rate is similar to the hit rate in sample and the values for all estimators ate identical with the exception of the case of zero specification error.

Table 2.5: Count of samples excluded from the Monte Carlo experiment[10]

J	K	N	Equal shares	One dominant	Half dominant	W=.6	W=.7	W=.8	W=.9	W=1
4	4	50	X							1
4	4	250	X							0
4	4	50		X				1	27	171
4	4	250		X						14
4	4	50			X				6	23
4	4	250			X					0
4	7	50	X					4	7	36
4	7	250	X							0
4	7	50		X			5	27	125	433
4	7	250		X						38
4	7	50			X			21	73	232
4	7	250			X					4
4	10	50	X							12
4	10	250	X							0
4	10	50		X		2	20	151	624	2468
4	10	250		X		-	-	-	-	-
4	10	50			X	4	30	165	684	1625
4	10	250			X			1	53	696
7	4	50	X						5	15
7	4	250	X							0
7	4	50		X			24	499	4797	70858
7	4	250		X		-	-	-	-	-
7	4	50			X	1	29	319	1722	6534
7	4	250			X			2	176	4919
7	7	50	X					1	9	44
7	7	250	X							1
7	7	50		X		1	45	591	4429	45836
7	7	250		X		-	-	-	-	-
7	7	50			X		34	264	1456	5339
7	7	250			X	-	-	-	-	-
7	10	50	X				3	33	91	256
7	10	250	X					1	10	250
7	10	50		X		6	102	798	4398	35046
7	10	250		X		-	-	-	-	-
7	10	50			X	-	-	-	-	-
7	10	250			X	-	-	-	-	-
10	4	50	X						1	27
10	4	250	X							0
10	4	50		X			3	106	1241	9669
10	4	250		X					167	35292
10	4	50			X		3	73	438	1958
10	4	250			X				20	1215
10	7	50	X				1	11	69	205
10	7	250	X						12	330
10	7	50		X		-	-	-	-	-
10	7	250		X		-	-	-	-	-
10	7	50			X	-	-	-	-	-
10	7	250			X	-	-	-	-	-
10	10	50	X			2	18	151	640	1816
10	10	250	X					1	234	3058
10	10	50		X		-	-	-	-	-
10	10	250		X		-	-	-	-	-
10	10	50			X	-	-	-	-	-
10	10	250			X	-	-	-	-	-

[10] The shaded areas indicate cases where more than 10% of the samples would be excluded from the simulation. The programs for these models were executed without a check for values of Estrella because of the high computation time.

2.5 Conclusions

The objective of this study is to explore the properties of Stein-rule estimators in the context of the orthonormal conditional logit model. We look at three different types of models: a model where all alternatives are equally likely, a model where one alternative is dominant, and a model where half of the alternatives are dominant.

The results of the Monte Carlo experiment show that Stein rule estimators have lower risk than the MLE both in terms of estimation and in terms of prediction for the entire parameter space under consideration. The risk improvement is larger in small samples and for small degrees of specification error. In addition, Stein rule performance improves relative to the MLE when the number of variables increases. In large samples the performance of all estimators improves because both the bias and the variance of the estimators decrease with an increase in the number of observations. The results for relative risk show no significant differences between the three cases.

When we increase the number of alternatives we confirm out previous results. Although higher number of alternatives generally does not favor shrinkage in estimation, the performance of all estimators improves. In out-of-sample prediction, higher number of alternatives expands the parameters space for which Stein rule estimators offer risk improvement over MLE in small samples.

In terms of recommendations, Stein rule should be used instead of the MLE in small samples and when the restrictions agree with the data. If we are uncertain of the quality of prior information or we have a large number of observations we can still use shrinkage estimation because it offers lower or equal risk relative to the maximum likelihood estimator over the entire space.

3 Risk Properties of a Stein-Like Estimator: Extensions of the Orthonormal Conditional Logit Model

3.1 Introduction

Chapter 2 explored the risk properties of Stein-like estimation in the context of the conditional logit model when the explanatory variables are orthonormal. This paper extends our analysis to a more general design matrix, allowing multicollinearity among the regressors.

The chapter is organized as follows. In section 2 we provide some theoretical background about multicollinearity in linear and nonlinear models, and explain how multicollinearity is modeled in the Monte Carlo experiment. Section 3 describes the shrinkage estimators and the Monte Carlo design. Section 4 contains our empirical results, and section 5 concludes.

3.2 Multicollinearity

3.2.1 Multicollinearity in the Linear Regression Model[11]

Multicollinearity is associated with the fact that economists or marketing researchers observe but do not control the values of the explanatory variables that produce or condition the values of the dependent variables. If multicollinearity is present, the statistical results are ambiguous because of interrelationships among the explanatory variables. In this case the variation in the dependent variable cannot be accurately attributed to a specific explanatory variable. As a result, the estimated coefficients tend to have large sampling error and thus the actual estimates may be far from the true parameter values. The negative consequences of multicollinearity in linear

[11] The discussion in the following two sections is heavily borrowed by Hill (1987).

regression are that the estimated coefficients may have incorrect signs and magnitudes, and the variables may not appear significantly different from zero, despite high R^2 or F values, indicating that a model fits the data well. In addition, the estimated coefficients may be sensitive to the addition or deletion of a few observations, or to the exclusion of a seemingly insignificant variable from the model.

To illustrate the collinearity problem we will consider a linear regression model

$$y = Z\beta + e, \tag{3.1}$$

where y is a vector of observations, Z is a matrix of explanatory variables, β is a vector of unknown regression coefficients, and $e \sim iid\ N(0, \sigma^2)$ is the vector of random disturbances.

Exact or perfect collinearity is said to exist when one of the variables can be written as an exact linear combination of the rest, and thus the columns of the matrix of explanatory variables are linearly dependent. Perfect collinearity does not occur very often in practice. Usually the linear relationship between the explanatory variables is not exact, but nearly exact, and has the form

$$Zc = z_1 c_1 + z_2 c_2 + ... + z_K c_K \doteq 0, \tag{3.2}$$

where c is a constant, and $\doteq$ means "almost equal to". The above expression can be written in the form of an auxiliary regression as

$$z_1 = z_2 d_2 + z_3 d_3 + ... + z_K d_K + \varepsilon_1, \tag{3.3}$$

where $d_i = -c_i / c_1$ and $\varepsilon_1 \sim N(0, \sigma_1^2)$. The smaller the value of σ_1^2, the stronger the linear dependence between the explanatory variables, and the more severe the multicollinearity. If $\sigma_1^2 = 0$, the multicollinearity would be perfect.

A more general framework of the collinearity problem can be provided by transforming the model in equation (3.1) into a principal components form. This transformation is based on the characteristic vectors of the $Z'Z$ matrix. If P is the orthogonal matrix whose columns are the characteristic vectors of $Z'Z$ then the equation (3.1) can be rewritten as

$$y = Z\beta + e = (ZP)(P'\beta) + e = M\theta + e . \tag{3.4}$$

The matrix M is called the matrix of principal components. The principal components have the property $m_i'm_i = \lambda_i$ and $m_i'm_j = 0$, where m_i is the i^{th} column of M and λ_i is the i^{th} characteristic root of $Z'Z$. Therefore $M'M = \Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_K)$ and the characteristic roots by convention are in descending order. In the transformed model collinearity is revealed by the magnitude of the characteristic roots of $Z'Z$. The existence of some relatively small characteristic roots implies the existence of some near-exact linear dependency among the columns of Z, and for every $\lambda_i = 0$ the linear dependency is exact.

Another transformation related to the principal components transformation is the singular value decomposition. The matrix of regressors can be written as

$$Z = U\Lambda^{1/2}P' , \tag{3.5}$$

where U is the matrix whose columns are the characteristic vectors of the matrix $Z'Z$, $\Lambda^{1/2}$ is the matrix containing the square roots of the characteristic roots of $Z'Z$ on the diagonal, and P is the matrix whose columns are the orthonormal characteristic vectors of $Z'Z$. The characteristic roots in this case are called the singular values of the matrix Z.

3.2.2 Multicollinearity in Nonlinear Models

Nonlinear statistical models can also suffer from ill-conditioned or multicollinear data. In the context of nonlinear models, studying the multicollinearity problem involves exploring the relationship between near-exact linear dependencies among a set of explanatory variables and the asymptotic covariance matrix of the estimator of the model's parameters. Studying the effects of multicollinearity on the small sample performance of the estimators is difficult to do analytically and often the problem is explored via Monte Carlo methods. In nonlinear models estimated by maximum likelihood methods, like the conditional and multinomial logit models, the asymptotic properties of the estimator can be approximated as

$$\hat{\beta} \sim N\left(\beta, \left[I(\beta)\right]^{-1}\right), \tag{3.6}$$

where the information matrix $I(\beta) = -E\dfrac{\partial^2 L}{\partial\beta\partial\beta'}$ is evaluated at $\hat{\beta}$. The effects of multicollinearity affect the ML estimator through $\left[I\left(\hat{\beta}\right)\right]^{-1}$. There are three popular nonlinear optimization procedures used to obtain ML estimates, and each of them used a different estimator of $I(\beta)$, which further complicates the problem. We may perceive the effects of multicollinearity differently depending on which optimization procedure we adopt, and also depending on the parameter values β. Despite the potential difficulties, the methods used to generate collinear data in the linear model are still applicable in the nonlinear case.

3.2.3 Modeling Multicollinearity in the Monte Carlo Experiment

In our study we analyze two types of multicollinearity. First we estimate models where collinearity is present among the explanatory variables related to each alternative. For example, in a marketing context, we may want to estimate the probability of choosing a particular brand of product as a function of its price, advertising expenditure, and quality, which could be measured by the number of attributes that the product offers. Let's consider a model about a choice between four competing brands. If Brand A offers a product of high quality, it is logical to expect that it is associated with high advertising expenditure and as a results charges a high price, which implies a positive relationship between the explanatory variables. Similar judgments may or may not be true for the other brands as well. Therefore, it is possible in this case collinearity to be present for brand A, but not for brands B, C or D, or it is possible to be present for more than one or all brands under consideration. We do not explore all possibilities and assume that collinearity is present for all brands.

Multicollinear data in our experiment are constructed in the following way. We start by creating a *K*x*K* matrix Σ which represents the desired correlation between the *K* explanatory variables for a given alternative. For example, if *K*=4 and we want to generate severe collinearity between two variables, the matrix of correlation is

$$\Sigma = \begin{bmatrix} 1 & .9 & 0 & 0 \\ .9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{3.7}$$

There exists a matrix C such that $CC' = \Sigma$. The matrix C can be written as $C = V\Lambda^{1/2}$, where V is a KxK matrix with columns equal to the characteristic vectors of

Σ, and $\Lambda^{1/2} = diag\left(\lambda_1^{1/2},...,\lambda_K^{1/2}\right)$ is a *K*x*K* diagonal matrix with elements equal to the square roots of the characteristic values of Σ. Let $z_i = Cu_i$, where u_i is a *K*x1 vector of independent standard normal random variables. Then $z_i \sim N(0,\Sigma)$ represents one observation for the *j*th alternative. Repeating this process *N* times generates a sample with the specified collinearity pattern. We generate a separate data matrix Z_i for each alternative, and then stack them horizontally to create the *N*x(*J*x*K*) matrix of explanatory variables *Z*, where *J* is the number of alternatives.

Secondly, we analyze the performance of the estimators when collinearity is present between alternative-specific variables. For example, consider a model that analyzes the choice of competing brands as a function of price and other product attributes. Price can be determined as a function of costs of production, and it is logical that similar products have similar costs. Likewise, prices may depend on the cost of advertising, and we expect similar products that use advertising as a way to compete to have similar advertising expenditures. Product prices may also be the result of assumptions about the consumers' price elasticity of demand. Since usually similar products compete for the same target market, we expect the prices of different brands to be influenced by the same factors that affect consumer demand. Finally, if prices are used as a tool of competition, usually a decrease in the price of one brand is followed by a decrease in the prices of the remaining brands.

Collinearity between alternatives was generated with the use of auxiliary regressions of the form presented in equation (3.3). We create an *N*x(*J*x*K*) matrix of explanatory variables *Z*, that are iid standard normal variables. To generate multi-

collinearity between the first two alternatives, we transformed the values of the second alternative as

$$z_2 = z_1 + cv_1, \tag{3.8}$$

where z_i is the relevant variable for each alternative, $v_1 \sim N(0,1)$ is a vector of random disturbances, and c is a constant that helps us achieve the desired degree of correlation between variables. The smaller the value of cv_1, the greater the linear dependence between z_1 and z_2, and the more severe the collinearity. Modeling collinearity between more alternatives was achieved in the same manner, where we generated a new vector of disturbances v_i for each separate case.

3.3 Design of Monte Carlo Experiment

The Monte Carlo experiment follows closely the design adopted in the case of orthonormal explanatory variables. The Stein rule estimator has the form

$$\delta = \left(1 - \frac{c}{u}\right)\beta_U + \left(\frac{c}{u}\right)\beta_R, \tag{3.9}$$

where β_U is the MLE (unrestricted) and β_R is the restricted MLE of β in the conditional logit model, c is a constant controlling the degree of shrinkage, and u is the value of the likelihood-ratio test statistic for the hypothesis $H_0 : \beta = 0$. The choice of shrinkage constant and test statistic is discussed in detail in the orthonormal case. The corresponding positive counterpart of the estimator is given by

$$\delta^+ = \left[1 - \frac{c}{u}\right]_+ \beta_U + \left(\frac{c}{u}\right)\beta_R, \tag{3.10}$$

where $[\arg]_+$ is a function that chooses the maximum of the argument or zero, ensuring that δ^+ is a convex combination or the restricted and the unrestricted maximum likelihood estimators.

Like in the orthonormal case, we use four different Stein-rule estimators, and their positive counterparts, based on the degree of shrinkage. We achieve this by choosing values of the shrinkage constant c from the interval $\{0.5, 1, 1.5 \text{ and } 2\}$, where higher values imply more shrinkage towards the restricted model.

The names of the estimators are presented below:

$$\begin{aligned} c_1 &= 0.5c: \text{ Stein05 and Pstein05} \\ c_2 &= c: \quad \text{Stein10 and Pstein10} \\ c_3 &= 1.5c: \text{ Stein15 and Pstein15} \\ c_4 &= 2c: \quad \text{Stein20 and Pstein20} \end{aligned} \tag{3.11}$$

In the Monte Carlo experiments we generate 2000 Monte Carlo samples for various degrees of specification error and other conditions as follows:

- We create a matrix of independent standard normal random variables. Then the variables are transformed using singular value decomposition to model collinearity between variables, or using auxiliary regressions to model collinearity between alternatives. The process of transforming the data is described in detail in section 3.2.3. Only one design matrix is generated per Monte Carlo experiment, and it remains fixed for each Monte Carlo sample.
- The "true" parameter vector β, used in the data generation process, is obtained as $\beta = w_i\beta_U + (1-w_i)\beta_R$, where β_U is a Kx1 vector with elements (1,1,…,1), β_R is a Kx1 vector of zeros, and $w_i = 0, 0.1, 0.2,\ldots,1$ controls the degree of specification error.

- For each Monte Carlo observation, the utility of individual *i* from alternative *j* is created as $U_{ij} = z'_{ij}\beta + e_{ij}$, where e_{ij} follows an extreme value (Gumbel) distribution. The observed variable y_{ij} is assigned a value of 1 if $U_{ij} = \max U_{ij}$, $j = 1,...,J$, and zero otherwise.
- For out-of-sample prediction, a holdout sample with N_O = 100 observations is generated as described above.
- In the Monte Carlo experiment, we manipulate the following:
    - number of variables: K = 4, 7, 10;
    - number of alternatives: J = 4, 7, 10;
    - sample size: N = 50, 250;
    - correlations between the explanatory variables. For each number of variables we explore two cases: severe collinearity (0.9) and low collinearity (0.4). For each case we generate collinearity between two variables and between all the variables for a given alternative;
    - correlations between alternatives. For one of the explanatory variables we generate severe collinearity (correlation $\doteq$ 0.9) and low collinearity (correlation $\doteq$ 0.4) between 2, 3 and 4 alternatives for J=4, 7 and 10 respectively.
    - mean of the explanatory variables, in order to create three cases: (1) all alternatives have similar shares, (2) one alternative is dominant, (3) half of the alternatives are dominant;
- For each estimator and each value of β the following estimates are obtained:
    - goodness of fit measures and information criteria for model selection;

- squared and weighted risk for MLE and Stein rule estimators;
- mean squared errors of prediction and hit rate out-of sample;

The goodness of fit measures and the different loss functions are described in detail in the orthonormal case.

The numerical results and the plots of the risk functions are presented in the appendix. The information in the tables contains the actual risk of the MLE and the relative risk of the Stein rule estimators, where values less than one indicate risk improvement over MLE.

3.4 Empirical Results

3.4.1 Collinearity Among Variables

3.4.1.1 Four Equally Likely Alternatives

- Goodness of Fit

Tables 3.1 and 3.2 show the goodness of fit measures when severe collinearity is present between two and among all explanatory variables, respectively. Tables 3.3 and 3.4 show the corresponding values in the case of low collinearity. All models of severe collinearity were estimated excluding samples with values of Estrella-R^2 higher than .98. Therefore, in these models the goodness of fit measures may be underestimated, but they can be used for comparison between models. The number of excluded samples from these models is presented in Table 3.13. The obtained values show that all models fit the data very well, and fit improves as we increase the signal-to-noise ratio. The values of Estrella-R^2 are higher for models with collinearity among all variables.

Table 3.1: Goodness of Fit Measures, Severe Collinearity Between Two Variables

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.07	-0.03	0.04	0.12	0.21	0.30	0.38	0.46	0.53	0.58
			BIC	3.00	2.99	2.95	2.88	2.80	2.69	2.58	2.47	2.36	2.25	2.16
			LR	4.11	4.77	6.77	10.31	14.52	19.78	25.06	30.92	36.07	41.82	46.20
4	4	250	Estrella	-0.02	0.00	0.05	0.12	0.20	0.30	0.39	0.47	0.55	0.62	0.68
			BIC	2.84	2.83	2.78	2.71	2.61	2.50	2.37	2.25	2.13	2.01	1.89
			LR	4.00	7.96	19.71	38.30	62.70	90.95	121.68	152.37	183.29	213.46	242.63
4	7	50	Estrella	-0.14	-0.11	0.00	0.13	0.27	0.41	0.52	0.61	0.68	0.74	0.78
			BIC	3.17	3.14	3.04	2.90	2.73	2.55	2.38	2.23	2.08	1.94	1.84
			LR	7.30	8.91	14.05	21.18	29.52	38.49	47.02	54.60	61.88	68.84	74.11
4	7	250	Estrella	-0.03	0.01	0.11	0.24	0.38	0.50	0.60	0.69	0.75	0.80	0.84
			BIC	2.90	2.86	2.76	2.61	2.43	2.25	2.08	1.92	1.77	1.65	1.53
			LR	7.02	15.94	41.95	80.11	124.15	169.48	211.82	252.71	288.11	318.71	348.06
4	10	50	Estrella	-0.21	-0.15	0.00	0.18	0.35	0.50	0.60	0.68	0.74	0.79	0.83
			BIC	3.35	3.30	3.15	2.95	2.74	2.52	2.36	2.20	2.05	1.93	1.82
			LR	10.46	12.90	20.48	30.20	40.96	51.52	59.70	67.75	75.04	81.26	86.86
4	10	250	Estrella	-0.04	0.02	0.17	0.34	0.50	0.62	0.72	0.79	0.84	0.87	0.90
			BIC	2.95	2.89	2.74	2.52	2.30	2.09	1.89	1.72	1.58	1.45	1.34
			LR	9.94	24.72	64.57	117.52	173.16	226.01	274.74	317.12	354.30	385.40	412.23
4	4	50	Estrella	-0.08	-0.07	-0.02	0.08	0.22	0.39	0.58	0.73	0.85	0.92	0.95
One dominant			BIC	3.00	2.99	2.94	2.84	2.68	2.46	2.17	1.85	1.53	1.24	1.05
			LR	4.10	4.78	7.38	12.35	20.14	31.31	45.91	61.82	77.85	92.04	101.84
4	4	250	Estrella	-0.02	0.00	0.06	0.16	0.30	0.46	0.62	0.76	0.86	0.93	0.97
One dominant			BIC	2.84	2.83	2.77	2.66	2.50	2.28	2.01	1.71	1.40	1.11	0.86
			LR	4.01	8.29	22.67	49.13	90.65	146.36	213.29	287.49	364.89	438.35	500.08
4	7	50	Estrella	-0.14	-0.11	0.02	0.20	0.40	0.59	0.74	0.85	0.91	0.94	0.96
One dominant			BIC	3.17	3.14	3.01	2.82	2.56	2.25	1.95	1.64	1.39	1.22	1.11
			LR	7.30	8.99	15.30	25.11	38.14	53.32	68.64	83.90	96.26	104.86	110.53
4	7	250	Estrella	-0.03	0.01	0.12	0.28	0.45	0.62	0.75	0.85	0.91	0.95	0.97
One dominant			BIC	2.90	2.86	2.75	2.57	2.33	2.06	1.78	1.49	1.23	1.00	0.85
			LR	6.93	16.27	44.65	90.39	149.75	217.22	286.78	358.38	423.73	481.93	519.95
4	10	50	Estrella	-0.21	-0.14	0.08	0.38	0.66	0.85	0.94	0.96	0.97	0.97	0.98
One dominant			BIC	3.35	3.29	3.07	2.70	2.22	1.73	1.38	1.22	1.15	1.12	1.10
			LR	10.45	13.42	24.32	42.92	66.59	91.37	108.92	116.65	120.19	121.93	122.99
4	10	250	Estrella	-0.04	0.02	0.19	0.40	0.60	0.76	0.86	0.93	0.96	0.98	0.98
One dominant			BIC	2.95	2.89	2.71	2.45	2.13	1.80	1.49	1.22	0.98	0.87	0.84
			LR	9.95	25.39	70.29	136.95	216.04	297.59	375.09	444.17	502.70	531.59	538.45
4	4	50	Estrella	-0.08	-0.07	-0.01	0.09	0.22	0.38	0.54	0.67	0.77	0.84	0.89
Two dominant			BIC	3.00	2.99	2.94	2.83	2.68	2.47	2.23	1.98	1.76	1.55	1.38
			LR	4.09	4.88	7.51	12.63	20.51	30.89	42.81	55.03	66.33	76.88	85.08
4	4	250	Estrella	-0.02	0.00	0.06	0.15	0.28	0.43	0.58	0.70	0.80	0.86	0.91
Two dominant			BIC	2.84	2.83	2.77	2.67	2.52	2.32	2.08	1.84	1.61	1.40	1.23
			LR	4.00	8.20	22.13	47.88	86.15	136.28	194.82	254.65	312.78	364.62	408.97
4	7	50	Estrella	-0.14	-0.10	0.03	0.21	0.41	0.60	0.72	0.82	0.87	0.91	0.93
Two dominant			BIC	3.17	3.14	3.01	2.80	2.54	2.24	1.99	1.74	1.56	1.41	1.30
			LR	7.31	9.09	15.70	25.96	39.00	53.94	66.76	79.01	88.08	95.56	101.12
4	7	250	Estrella	-0.03	0.01	0.12	0.28	0.45	0.61	0.74	0.83	0.89	0.93	0.95
Two dominant			BIC	2.90	2.86	2.75	2.56	2.32	2.06	1.80	1.55	1.34	1.17	1.02
			LR	6.92	16.34	45.24	92.07	150.79	216.16	282.44	343.89	396.17	440.53	476.68
4	10	50	Estrella	-0.21	-0.15	0.05	0.29	0.52	0.69	0.81	0.87	0.91	0.93	0.95
Two dominant			BIC	3.35	3.29	3.10	2.82	2.49	2.16	1.88	1.66	1.50	1.39	1.32
			LR	10.45	13.19	22.75	36.90	53.11	69.60	83.64	94.59	102.60	108.04	111.75
4	10	250	Estrella	-0.04	0.02	0.20	0.42	0.63	0.77	0.87	0.92	0.95	0.97	0.97
Two dominant			BIC	2.95	2.89	2.70	2.41	2.08	1.76	1.48	1.25	1.07	0.95	0.88
			LR	9.94	25.94	73.63	145.22	227.61	307.72	378.14	436.00	480.68	512.10	528.76

Table 3.2: Goodness of Fit Measures, Severe Collinearity Among All Variables[12]

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.05	0.03	0.16	0.28	0.40	0.50	0.59	0.66	0.72	0.76
			BIC	3.00	2.97	2.89	2.75	2.60	2.44	2.30	2.14	2.01	1.89	1.78
			LR	4.11	5.67	9.84	16.82	24.24	32.08	39.37	47.19	53.81	60.02	65.26
4	4	250	Estrella	-0.02	0.01	0.10	0.21	0.34	0.46	0.57	0.66	0.73	0.78	0.83
			BIC	2.84	2.81	2.73	2.60	2.44	2.27	2.10	1.94	1.79	1.65	1.53
			LR	4.00	11.71	33.44	66.15	105.35	148.57	189.89	230.69	267.83	302.52	333.67
4	7	50	Estrella	-0.14	0.03	0.33	0.58	0.74	0.83	0.88	0.92	0.94	0.95	0.96
			BIC	3.17	3.01	2.65	2.27	1.95	1.70	1.51	1.35	1.25	1.18	1.12
			LR	7.30	15.59	33.53	52.58	68.27	80.84	90.73	98.36	103.54	107.11	109.86
4	7	250	Estrella	-0.03	0.12	0.39	0.61	0.76	0.84	0.89	0.93	0.95	0.96	0.97
			BIC	2.90	2.75	2.42	2.06	1.76	1.52	1.32	1.17	1.04	0.95	0.88
			LR	7.02	44.70	127.94	215.55	291.52	351.43	400.86	439.96	471.89	495.04	512.05
4	10	50	Estrella	-0.21	0.12	0.55	0.77	0.87	0.92	0.94	0.95	0.96	0.96	0.97
			BIC	3.35	3.02	2.44	1.99	1.68	1.49	1.36	1.29	1.24	1.21	1.18
			LR	10.46	26.93	55.64	78.44	93.72	103.23	109.56	113.41	115.82	117.48	118.67
4	10	250	Estrella	-0.04	0.31	0.68	0.85	0.92	0.95	0.97	0.97	0.98	0.98	0.98
			BIC	2.95	2.57	1.97	1.53	1.24	1.05	0.93	0.88	0.85	0.84	0.83
			LR	9.94	106.05	256.58	367.06	439.07	486.48	515.92	529.59	535.34	538.31	539.71
4	4	50	Estrella	-0.08	-0.05	0.04	0.19	0.36	0.52	0.65	0.77	0.84	0.90	0.94
One dominant			BIC	3.00	2.97	2.88	2.71	2.50	2.26	2.03	1.77	1.55	1.33	1.15
			LR	4.11	5.70	10.42	18.80	29.36	41.10	53.01	65.73	76.83	88.00	96.74
4	4	250	Estrella	-0.02	0.02	0.11	0.26	0.43	0.59	0.72	0.82	0.89	0.94	0.97
One dominant			BIC	2.84	2.81	2.71	2.54	2.32	2.07	1.81	1.54	1.29	1.07	0.87
			LR	4.00	12.01	37.54	79.18	134.68	197.40	263.54	330.78	392.78	448.28	498.95
4	7	50	Estrella	-0.14	0.03	0.37	0.63	0.80	0.88	0.93	0.95	0.96	0.97	0.97
One dominant			BIC	3.17	3.00	2.61	2.17	1.80	1.51	1.32	1.21	1.13	1.08	1.05
			LR	7.30	15.99	35.69	57.42	76.09	90.50	99.98	105.70	109.37	111.87	113.69
4	7	250	Estrella	-0.03	0.12	0.40	0.64	0.79	0.88	0.93	0.96	0.97	0.98	0.98
One dominant			BIC	2.90	2.75	2.40	2.02	1.68	1.39	1.17	0.98	0.86	0.82	0.80
			LR	6.92	45.52	132.68	226.75	312.26	383.60	440.01	487.32	517.19	528.04	532.49
4	10	50	Estrella	-0.21	0.13	0.56	0.78	0.88	0.92	0.94	0.96	0.96	0.97	0.97
One dominant			BIC	3.35	3.01	2.42	1.95	1.66	1.47	1.35	1.27	1.22	1.19	1.16
			LR	10.46	27.33	56.91	80.05	94.88	104.20	110.17	114.28	116.70	118.41	119.60
4	10	250	Estrella	-0.04	0.31	0.69	0.86	0.93	0.96	0.97	0.98	0.99	0.99	0.99
One dominant			BIC	2.95	2.57	1.96	1.51	1.22	1.01	0.87	0.78	0.71	0.67	0.64
			LR	9.95	105.98	257.54	369.93	443.95	494.73	530.41	553.78	570.20	581.56	589.4
4	4	50	Estrella	-0.08	-0.05	0.04	0.18	0.33	0.48	0.61	0.73	0.81	0.87	0.91
Two dominant			BIC	3.00	2.97	2.88	2.73	2.54	2.32	2.10	1.86	1.66	1.47	1.29
			LR	4.11	5.68	10.14	17.96	27.41	38.25	49.14	61.03	71.39	81.01	89.66
4	4	250	Estrella	-0.02	0.02	0.11	0.26	0.42	0.58	0.70	0.80	0.87	0.91	0.94
Two dominant			BIC	2.84	2.81	2.71	2.55	2.33	2.09	1.84	1.61	1.40	1.20	1.04
			LR	4.00	12.35	37.65	78.98	132.17	193.43	253.99	313.50	366.47	414.51	454.96
4	7	50	Estrella	-0.14	0.03	0.35	0.61	0.77	0.87	0.91	0.94	0.95	0.96	0.97
Two dominant			BIC	3.17	3.01	2.63	2.22	1.87	1.57	1.38	1.25	1.17	1.12	1.08
			LR	7.31	15.66	34.38	55.06	72.75	87.43	97.14	103.39	107.28	110.02	112.06
4	7	250	Estrella	-0.03	0.12	0.41	0.64	0.79	0.87	0.92	0.95	0.96	0.97	0.98
Two dominant			BIC	2.90	2.74	2.39	2.02	1.68	1.42	1.22	1.05	0.92	0.84	0.81
			LR	6.91	45.79	133.54	227.49	311.41	377.81	427.83	468.97	501.32	521.14	529.83
4	10	50	Estrella	-0.21	0.13	0.56	0.78	0.88	0.92	0.95	0.96	0.96	0.97	0.97
Two dominant			BIC	3.35	3.01	2.42	1.96	1.64	1.45	1.32	1.25	1.20	1.17	1.15
			LR	10.45	27.40	56.90	79.94	95.56	105.38	111.68	115.21	117.50	119.14	120.29
4	10	250	Estrella	-0.04	0.31	0.68	0.86	0.93	0.96	0.98	0.98	0.99	0.99	0.99
Two dominant			BIC	2.95	2.57	1.96	1.51	1.22	1.01	0.86	0.76	0.69	0.63	0.58
			LR	9.94	106.31	257.20	370.17	444.49	495.54	532.17	557.97	576.80	591.96	603.7

[12] The models with K=10, N=250 and one dominant or two dominant alternatives are estimated without a check for values of Estrella-R^2.

Table 3.3: Goodness of Fit Measures, Low Collinearity Between Two Variables[13]

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.06	0.00	0.09	0.20	0.31	0.41	0.51	0.58	0.65	0.70
			BIC	3.00	2.98	2.92	2.82	2.71	2.57	2.43	2.29	2.16	2.02	1.92
			LR	4.11	5.13	8.10	13.14	18.99	26.01	32.91	39.87	46.41	53.12	58.50
4	4	250	Estrella	-0.02	0.01	0.07	0.16	0.27	0.37	0.47	0.56	0.63	0.70	0.75
			BIC	2.84	2.82	2.76	2.66	2.53	2.40	2.26	2.12	1.98	1.86	1.74
			LR	4.00	9.51	25.76	50.08	81.62	116.31	150.47	186.16	219.62	250.59	279.96
4	7	50	Estrella	-0.14	-0.10	0.02	0.18	0.34	0.47	0.58	0.66	0.73	0.77	0.81
			BIC	3.17	3.14	3.01	2.84	2.65	2.45	2.27	2.12	1.98	1.86	1.76
			LR	7.30	9.20	15.46	24.01	33.70	43.61	52.48	60.26	67.04	72.77	78.22
4	7	250	Estrella	-0.03	0.02	0.13	0.28	0.43	0.56	0.65	0.73	0.79	0.83	0.86
			BIC	2.90	2.86	2.73	2.56	2.36	2.16	1.99	1.82	1.68	1.56	1.44
			LR	7.02	17.85	48.51	92.96	142.26	191.16	235.23	276.64	312.94	342.89	371.21
4	10	50	Estrella	-0.21	-0.14	0.03	0.22	0.40	0.55	0.65	0.73	0.78	0.83	0.86
			BIC	3.35	3.29	3.12	2.90	2.67	2.44	2.26	2.09	1.95	1.82	1.71
			LR	10.46	13.31	21.86	32.60	44.12	55.73	64.51	73.02	80.33	86.85	92.02
4	10	250	Estrella	-0.04	0.03	0.19	0.38	0.54	0.66	0.75	0.81	0.86	0.89	0.91
			BIC	2.95	2.89	2.71	2.48	2.24	2.02	1.81	1.65	1.50	1.38	1.28
			LR	9.94	26.64	71.26	129.41	188.87	244.36	294.88	336.38	372.78	403.14	428.57
4	4	50	Estrella	-0.08	-0.06	0.01	0.13	0.28	0.44	0.61	0.74	0.84	0.90	0.94
One dominant			BIC	3.00	2.98	2.91	2.79	2.61	2.38	2.11	1.83	1.56	1.31	1.12
			LR	4.10	5.13	8.61	14.94	23.73	35.05	48.60	62.60	76.48	88.72	98.42
4	4	250	Estrella	-0.02	0.01	0.08	0.20	0.35	0.51	0.66	0.78	0.87	0.93	0.96
One dominant			BIC	2.84	2.82	2.75	2.62	2.43	2.20	1.93	1.65	1.37	1.10	0.88
			LR	4.01	9.67	28.36	60.73	107.45	165.63	232.83	303.15	373.70	439.33	494.90
4	7	50	Estrella	-0.14	-0.10	0.06	0.27	0.49	0.68	0.82	0.90	0.94	0.96	0.97
One dominant			BIC	3.17	3.13	2.98	2.73	2.42	2.08	1.73	1.44	1.24	1.12	1.05
			LR	7.30	9.41	17.17	29.40	44.91	62.15	79.30	94.00	103.88	109.84	113.41
4	7	250	Estrella	-0.03	0.02	0.15	0.33	0.52	0.69	0.81	0.90	0.95	0.97	0.98
One dominant			BIC	2.90	2.85	2.72	2.50	2.22	1.92	1.61	1.31	1.05	0.86	0.80
			LR	6.93	18.26	52.71	107.33	176.52	252.38	328.22	403.84	470.46	517.85	531.50
4	10	50	Estrella	-0.21	-0.14	0.06	0.31	0.55	0.73	0.85	0.92	0.96	0.98	0.99
One dominant			BIC	3.35	3.28	3.08	2.79	2.44	2.07	1.75	1.46	1.20	0.97	0.85
			LR	10.45	13.53	23.73	38.46	55.66	74.29	90.47	104.62	117.82	129.01	135.40
4	10	250	Estrella	-0.04	0.03	0.21	0.43	0.63	0.77	0.87	0.93	0.96	0.98	0.99
One dominant			BIC	2.95	2.88	2.68	2.40	2.08	1.76	1.46	1.20	0.97	0.78	0.64
			LR	9.95	27.42	77.17	149.15	229.36	309.07	383.45	449.34	506.16	552.53	589.46
4	4	50	Estrella	-0.08	-0.06	0.01	0.13	0.27	0.43	0.57	0.69	0.79	0.85	0.90
Two dominant			BIC	3.00	2.98	2.91	2.79	2.61	2.41	2.17	1.94	1.72	1.51	1.34
			LR	4.09	5.20	8.64	14.97	23.54	33.99	45.54	57.13	68.47	78.94	87.28
4	4	250	Estrella	-0.02	0.01	0.08	0.19	0.34	0.49	0.63	0.74	0.83	0.88	0.92
Two dominant			BIC	2.84	2.82	2.75	2.62	2.44	2.22	1.99	1.75	1.53	1.33	1.16
			LR	4.00	9.83	28.38	59.83	104.32	159.00	217.42	278.40	333.20	382.37	425.43
4	7	50	Estrella	-0.14	-0.10	0.06	0.28	0.49	0.67	0.79	0.86	0.90	0.93	0.94
Two dominant			BIC	3.17	3.13	2.97	2.72	2.42	2.09	1.83	1.59	1.43	1.32	1.23
			LR	7.31	9.50	17.53	30.00	45.03	61.40	74.69	86.46	94.38	100.22	104.31
4	7	250	Estrella	-0.03	0.02	0.15	0.33	0.52	0.68	0.79	0.87	0.92	0.95	0.97
Two dominant			BIC	2.90	2.85	2.71	2.49	2.22	1.94	1.66	1.41	1.21	1.04	0.92
			LR	7.08	18.36	53.45	108.58	175.90	247.98	316.95	379.55	430.52	472.17	502.81
4	10	50	Estrella	-0.21	-0.14	0.07	0.32	0.54	0.70	0.81	0.87	0.91	0.94	0.95
Two dominant			BIC	3.35	3.28	3.07	2.78	2.46	2.14	1.88	1.67	1.50	1.36	1.26
			LR	10.45	13.61	24.07	38.62	54.78	70.66	83.96	94.43	102.87	109.93	114.65
4	10	250	Estrella	-0.04	0.03	0.22	0.46	0.65	0.79	0.88	0.93	0.95	0.97	0.98
Two dominant			BIC	2.95	2.88	2.67	2.37	2.03	1.71	1.44	1.22	1.05	0.92	0.81
			LR	9.94	27.98	80.67	157.02	240.85	320.15	387.62	442.98	486.38	518.67	545.00

[13] The models with 10 variables are estimated without a check for values of Estrella-R^2.

Table 3.4: Goodness of Fit Measures, Low Collinearity Among All Variables[14]

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.05	0.04	0.17	0.30	0.43	0.53	0.63	0.70	0.76	0.80
			BIC	3.00	2.97	2.88	2.74	2.58	2.41	2.24	2.08	1.93	1.79	1.67
			LR	4.11	5.69	10.17	17.32	25.43	33.88	42.19	50.49	57.84	64.74	70.73
4	4	250	Estrella	-0.02	0.01	0.10	0.22	0.34	0.46	0.56	0.64	0.71	0.77	0.81
			BIC	2.84	2.81	2.73	2.59	2.44	2.28	2.12	1.96	1.82	1.69	1.58
			LR	4.00	11.79	33.63	66.68	104.53	146.21	185.40	224.72	259.45	291.84	320.42
4	7	50	Estrella	-0.14	-0.03	0.21	0.45	0.62	0.75	0.82	0.87	0.91	0.93	0.94
			BIC	3.17	3.07	2.80	2.49	2.20	1.93	1.72	1.55	1.42	1.32	1.24
			LR	7.30	12.71	25.89	41.69	56.10	69.53	79.91	88.58	95.16	99.97	104.03
4	7	250	Estrella	-0.03	0.07	0.28	0.50	0.65	0.76	0.83	0.88	0.91	0.93	0.95
			BIC	2.90	2.80	2.56	2.26	1.98	1.74	1.54	1.38	1.25	1.13	1.03
			LR	7.02	31.43	92.33	167.04	235.77	295.96	345.57	387.78	420.38	449.46	474.91
4	10	50	Estrella	-0.21	-0.02	0.32	0.59	0.75	0.84	0.89	0.93	0.95	0.96	0.97
			BIC	3.35	3.16	2.78	2.37	2.04	1.78	1.58	1.42	1.30	1.19	1.11
			LR	10.46	19.57	38.82	59.05	75.89	88.62	98.61	106.77	112.83	118.31	122.33
4	10	250	Estrella	-0.04	0.17	0.50	0.72	0.84	0.90	0.94	0.96	0.97	0.98	0.98
			BIC	2.95	2.73	2.29	1.88	1.56	1.32	1.15	1.02	0.92	0.84	0.77
			LR	9.94	65.17	175.66	278.94	358.91	417.13	460.69	493.95	518.84	538.45	555.62
4	4	50	Estrella	-0.08	-0.05	0.05	0.21	0.38	0.54	0.66	0.77	0.85	0.90	0.94
One dominant			BIC	3.00	2.97	2.87	2.70	2.47	2.24	2.00	1.76	1.53	1.33	1.16
			LR	4.11	5.75	10.88	19.51	30.78	42.37	54.16	66.16	77.59	87.78	96.32
4	4	250	Estrella	-0.02	0.02	0.11	0.26	0.42	0.58	0.71	0.82	0.89	0.94	0.97
One dominant			BIC	2.84	2.81	2.71	2.55	2.33	2.08	1.82	1.55	1.29	1.06	0.86
			LR	4.00	12.32	37.40	78.57	132.00	195.55	261.07	328.11	391.82	450.19	500.28
4	7	50	Estrella	-0.14	-0.03	0.23	0.49	0.68	0.79	0.86	0.90	0.93	0.95	0.96
One dominant			BIC	3.17	3.06	2.77	2.42	2.09	1.81	1.59	1.43	1.30	1.19	1.11
			LR	7.30	12.95	27.30	45.10	61.59	75.68	86.44	94.54	100.87	106.50	110.50
4	7	250	Estrella	-0.03	0.07	0.30	0.54	0.71	0.83	0.90	0.94	0.97	0.98	0.98
One dominant			BIC	2.90	2.80	2.53	2.20	1.87	1.57	1.31	1.09	0.92	0.83	0.80
			LR	7.16	32.33	98.10	182.15	265.23	340.24	404.49	460.23	502.59	525.00	531.80
4	10	50	Estrella	-0.21	-0.01	0.35	0.63	0.78	0.87	0.92	0.95	0.96	0.98	0.99
One dominant			BIC	3.35	3.16	2.74	2.29	1.95	1.67	1.47	1.30	1.15	1.00	0.90
			LR	10.45	19.96	40.84	63.11	80.23	94.34	104.29	112.62	120.14	127.78	132.78
4	10	250	Estrella	-0.04	0.17	0.51	0.73	0.85	0.91	0.95	0.97	0.98	0.99	0.99
One dominant			BIC	2.95	2.73	2.28	1.85	1.53	1.28	1.09	0.95	0.83	0.74	0.66
			LR	9.95	65.78	178.45	284.77	366.18	428.26	475.08	511.27	540.28	563.13	582.22
4	4	50	Estrella	-0.08	-0.05	0.05	0.19	0.36	0.50	0.64	0.74	0.82	0.88	0.91
Two dominant			BIC	3.00	2.97	2.87	2.71	2.50	2.29	2.06	1.84	1.63	1.43	1.26
			LR	4.10	5.71	10.65	18.67	29.26	39.86	51.34	62.53	73.02	82.61	91.05
4	4	250	Estrella	-0.02	0.02	0.12	0.27	0.43	0.58	0.71	0.80	0.87	0.91	0.94
Two dominant			BIC	2.84	2.81	2.71	2.54	2.32	2.08	1.83	1.60	1.38	1.20	1.05
			LR	3.99	12.53	38.48	81.27	135.62	196.48	257.70	315.77	369.54	416.18	453.83
4	7	50	Estrella	-0.14	-0.03	0.23	0.49	0.67	0.80	0.87	0.91	0.93	0.95	0.96
Two dominant			BIC	3.17	3.06	2.78	2.43	2.09	1.79	1.57	1.40	1.28	1.19	1.13
			LR	7.31	12.83	26.95	44.55	61.39	76.29	87.42	96.01	102.0	106.5	109.8
4	7	250	Estrella	-0.03	0.07	0.30	0.53	0.70	0.82	0.88	0.93	0.95	0.97	0.97
Two dominant			BIC	2.90	2.80	2.53	2.20	1.89	1.60	1.37	1.18	1.02	0.90	0.83
			LR	6.91	32.31	98.56	181.24	260.48	331.36	388.17	436.68	476.42	505.60	523.31
4	10	50	Estrella	-0.21	-0.01	0.35	0.62	0.78	0.86	0.91	0.94	0.96	0.98	0.98
Two dominant			BIC	3.35	3.16	2.74	2.31	1.96	1.69	1.50	1.33	1.19	1.05	0.95
			LR	10.45	19.89	40.58	62.10	79.71	93.06	102.94	111.06	118.14	125.31	130.06
4	10	250	Estrella	-0.04	0.17	0.51	0.73	0.85	0.91	0.95	0.96	0.98	0.98	0.99
Two dominant			BIC	2.95	2.73	2.28	1.86	1.53	1.28	1.10	0.96	0.86	0.77	0.70
			LR	9.94	65.42	177.64	283.26	365.11	427.19	473.17	507.58	534.47	555.16	572.96

[14] The models with 10 variables are estimated without a check for values of Estrella-R^2.

When collinearity is present in linear regression, the models often have very high values of R^2, but the coefficient may have wrong signs, unreasonable magnitudes, or may not be statistically significant from zero. Collinearity may have similar effects in the nonlinear case, which would explain the high goodness of fit values. Models with higher number of variables show better fit in all the four cases under consideration.

- Squared Error Loss

The squared error loss measures the distance of the estimator from the true parameter, thus the ability to estimate the parameters of the model. The squared error loss increases as the bias and/or the variance of the estimators increase. Multicollinearity increases the variance of the estimators, while shrinkage decreases variability at the expense of possible bias. Therefore exploring the performance of shrinkage estimators is interesting when collinearity is present among the regressors. If shrinkage improves estimation in choice models with collinear variables it may provide a better alternative to social scientists or market researchers who have to use "problematic" data to estimate their models.

In our experiment we generate four different types of collinear data: severe collinearity (correlation close to 0.9) between two variables and among all variables, and low collinearity (correlation close to 0.4) between two variables and among all variables for each alternative. Tables 3.5 and 3.6 show the relative risk values in terms of squared error loss for the two cases of severe collinearity. Because of the large volume of presented information and the similarity of the results, the values and risk functions for the case of low collinearity and the cases of one or two dominant alternatives are discussed but not included in the text.

Table 3.5: Squared Error Loss, Four Equally Likely Alternatives, Severe Collinearity Between Two Variables

	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	0.40	0.40	0.39	0.42	0.46	0.49	0.49	0.58	0.61	0.70	0.75
	PStein c=0.5	0.62	0.67	0.78	0.85	0.90	0.92	0.94	0.95	0.95	0.95	0.96
	PStein c=1	0.39	0.46	0.63	0.75	0.82	0.86	0.89	0.90	0.92	0.92	0.92
	PStein c=1.5	0.24	0.33	0.53	0.68	0.77	0.83	0.86	0.87	0.89	0.88	0.89
	PStein c=2	0.15	0.24	0.47	0.65	0.74	0.80	0.84	0.85	0.87	0.86	0.87
N=250	Risk MLE	0.07	0.07	0.07	0.08	0.08	0.09	0.09	0.10	0.10	0.11	0.13
	PStein c=0.5	0.61	0.79	0.92	0.96	0.98	0.98	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.37	0.66	0.86	0.93	0.96	0.97	0.98	0.98	0.98	0.98	0.98
	PStein c=1.5	0.22	0.58	0.82	0.91	0.95	0.96	0.97	0.97	0.98	0.98	0.97
	PStein c=2	0.14	0.53	0.81	0.90	0.94	0.95	0.96	0.97	0.97	0.97	0.97
						K=7						
N=50	Risk MLE	0.53	0.53	0.57	0.59	0.65	0.74	0.89	1.06	1.30	1.78	2.18
	PStein c=0.5	0.48	0.57	0.72	0.81	0.85	0.88	0.90	0.90	0.91	0.91	0.91
	PStein c=1	0.22	0.33	0.55	0.68	0.75	0.80	0.81	0.83	0.84	0.84	0.84
	PStein c=1.5	0.09	0.22	0.46	0.62	0.70	0.74	0.76	0.77	0.78	0.78	0.78
	PStein c=2	0.04	0.17	0.43	0.61	0.68	0.72	0.73	0.73	0.74	0.73	0.73
N=250	Risk MLE	0.08	0.08	0.09	0.10	0.10	0.11	0.13	0.15	0.17	0.18	0.21
	PStein c=0.5	0.46	0.76	0.91	0.95	0.96	0.97	0.98	0.98	0.98	0.98	0.98
	PStein c=1	0.20	0.63	0.85	0.92	0.94	0.95	0.96	0.96	0.96	0.96	0.96
	PStein c=1.5	0.08	0.59	0.83	0.90	0.93	0.93	0.95	0.94	0.95	0.95	0.94
	PStein c=2	0.03	0.60	0.84	0.89	0.92	0.93	0.94	0.94	0.94	0.94	0.93
						K=10						
N=50	Risk MLE	0.57	0.60	0.68	0.77	0.95	1.21	1.42	1.91	2.66	3.81	5.81
	PStein c=0.5	0.43	0.53	0.68	0.77	0.82	0.84	0.85	0.86	0.87	0.87	0.88
	PStein c=1	0.15	0.29	0.51	0.64	0.69	0.72	0.74	0.75	0.76	0.77	0.78
	PStein c=1.5	0.05	0.20	0.45	0.59	0.63	0.65	0.67	0.67	0.67	0.68	0.69
	PStein c=2	0.01	0.18	0.46	0.61	0.62	0.62	0.63	0.62	0.61	0.61	0.62
N=250	Risk MLE	0.10	0.11	0.12	0.12	0.14	0.16	0.20	0.22	0.26	0.30	0.36
	PStein c=0.5	0.40	0.75	0.90	0.94	0.96	0.96	0.97	0.96	0.96	0.96	0.96
	PStein c=1	0.13	0.61	0.84	0.90	0.92	0.94	0.94	0.94	0.93	0.93	0.93
	PStein c=1.5	0.03	0.57	0.81	0.88	0.91	0.92	0.92	0.92	0.91	0.91	0.91
	PStein c=2	0.01	0.61	0.82	0.89	0.90	0.92	0.91	0.91	0.90	0.89	0.89

Table 3.6: Squared Error Loss, Four Equally Likely Alternatives, Severe Collinearity Among All Variables[15]

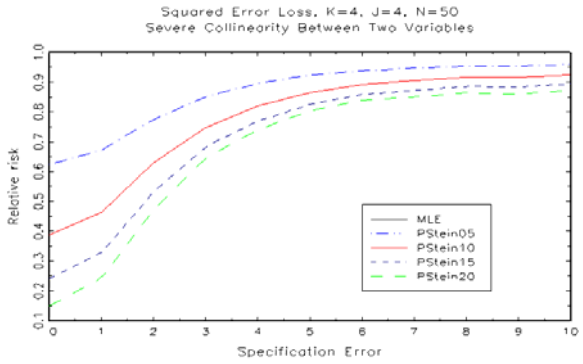
	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	0.93	0.91	0.93	1.03	1.11	1.23	1.31	1.44	1.60	1.74	1.91
	PStein c=0.5	0.62	0.69	0.81	0.89	0.92	0.94	0.95	0.96	0.96	0.97	0.97
	PStein c=1	0.38	0.48	0.66	0.79	0.85	0.89	0.91	0.92	0.93	0.93	0.94
	PStein c=1.5	0.24	0.34	0.54	0.70	0.78	0.83	0.86	0.88	0.90	0.90	0.91
	PStein c=2	0.15	0.24	0.45	0.63	0.72	0.79	0.82	0.85	0.87	0.87	0.88
N=250	Risk MLE	0.16	0.16	0.17	0.18	0.19	0.21	0.24	0.26	0.28	0.31	0.35
	PStein c=0.5	0.61	0.83	0.94	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.37	0.70	0.89	0.95	0.96	0.97	0.98	0.98	0.98	0.99	0.99
	PStein c=1.5	0.22	0.59	0.85	0.92	0.95	0.96	0.97	0.97	0.98	0.98	0.98
	PStein c=2	0.13	0.50	0.80	0.90	0.93	0.95	0.96	0.96	0.97	0.97	0.97
						K=7						
N=50	Risk MLE	1.96	2.15	2.63	3.40	4.60	6.18	7.98	10.24	*12.38*	*13.57*	*14.21*
	PStein c=0.5	0.49	0.71	0.86	0.91	0.93	0.94	0.95	0.95	*0.95*	*0.95*	*0.95*
	PStein c=1	0.22	0.49	0.73	0.82	0.86	0.88	0.89	0.90	*0.90*	*0.91*	*0.91*
	PStein c=1.5	0.09	0.32	0.62	0.74	0.80	0.83	0.84	0.85	*0.86*	*0.86*	*0.86*
	PStein c=2	0.04	0.21	0.52	0.67	0.74	0.77	0.79	0.81	*0.81*	*0.82*	*0.82*
N=250	Risk MLE	0.33	0.35	0.40	0.47	0.58	0.68	0.80	0.92	1.11	1.20	*1.32*
	PStein c=0.5	0.46	0.89	0.96	0.98	0.98	0.99	0.99	0.99	0.99	0.99	*0.99*
	PStein c=1	0.20	0.79	0.92	0.95	0.97	0.97	0.97	0.98	0.98	0.98	*0.98*
	PStein c=1.5	0.08	0.70	0.89	0.93	0.95	0.96	0.96	0.96	0.96	0.97	*0.97*
	PStein c=2	0.03	0.61	0.85	0.91	0.93	0.94	0.95	0.95	0.95	0.96	*0.96*
						K=10						
N=50	Risk MLE	3.07	3.53	5.60	9.40	20.0	41.3	*56.3*	*80.4*	*81.3*	*87.4*	*92.5*
	PStein c=0.5	0.43	0.73	0.87	0.91	0.92	0.93	*0.93*	*0.93*	*0.93*	*0.93*	*0.93*
	PStein c=1	0.15	0.51	0.75	0.82	0.85	0.86	*0.86*	*0.86*	*0.86*	*0.86*	*0.87*
	PStein c=1.5	0.05	0.34	0.64	0.73	0.77	0.79	*0.80*	*0.80*	*0.80*	*0.80*	*0.80*
	PStein c=2	0.01	0.22	0.54	0.65	0.71	0.73	*0.73*	*0.74*	*0.74*	*0.74*	*0.74*
N=250	Risk MLE	0.50	0.59	0.81	1.10	1.44	1.77	*2.08*	*2.02*	*2.03*	*2.03*	*2.29*
	PStein c=0.5	0.40	0.93	0.97	0.98	0.98	0.98	*0.98*	*0.99*	*0.99*	*1.00*	*1.00*
	PStein c=1	0.13	0.85	0.94	0.96	0.96	0.97	*0.97*	*0.97*	*0.98*	*0.99*	*1.01*
	PStein c=1.5	0.04	0.79	0.91	0.93	0.94	0.95	*0.95*	*0.96*	*0.97*	*0.99*	*1.01*
	PStein c=2	0.01	0.72	0.88	0.91	0.93	0.93	*0.94*	*0.95*	*0.96*	*0.99*	*1.01*

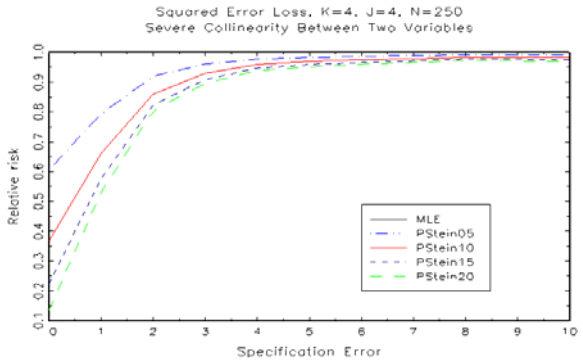
The results for relative risk in the case of severe collinearity show that shrinkage estimators dominate the maximum likelihood estimator over the entire parameter space.

[15] The bar indicates that for the corresponding weights to the right of it, more than 10% of the Monte Carlo samples were replaced. As a result, the values in italic cannot be considered representative of our Monte Carlo experiment.

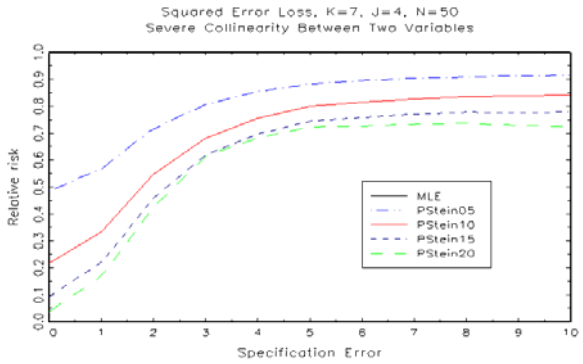
For N=50 and K=4 the risk differences are statistically significant for each level of specification error, with the exception of the risk of the minimum shrinkage estimator Pstein05 which gets close to the risk of MLE after the mid-range of parameter space. There is a clear distinction between the Stein rule estimators indicating that more shrinkage offers larger risk improvement, thus the dominant estimator is Pstein20. The risk improvement is substantial for values of Estrella-R^2 less than .4. As we increase the sample size the shrinkage estimators still have lower risk than the MLE for each degree of specification error, but their risks approach the risk of MLE for values of Estrella-R^2 less than .3. We expect faster convergence in large samples because the bias and variance of the maximum likelihood estimator decrease and the benefits of shrinking the estimates towards zero decrease as well. Also in this case more shrinkage offers larger risk improvement and Pstein20 is the dominant estimator. These results are confirmed also in the case of collinearity between all variables and there are no significant differences in the risk functions of the estimators. In terms of magnitude, the risk of all estimators is larger when collinearity is present among all variables, compared to the case of two variables, which is an expected result, because in this case the estimators are likely to have higher variance. When w=1 and N=50, the risk of MLE is about 155% higher when severe collinearity is present among all variables, compared to severe collinearity between two variables. Figure 3.1 shows the risk functions when severe collinearity is present between two variables, and Figure 3.2 shows the case of severe collinearity among all variables. In both cases all alternatives are equally likely. When one or two of the alternatives are dominant, the relative risk functions are very similar and therefore not included in the text.

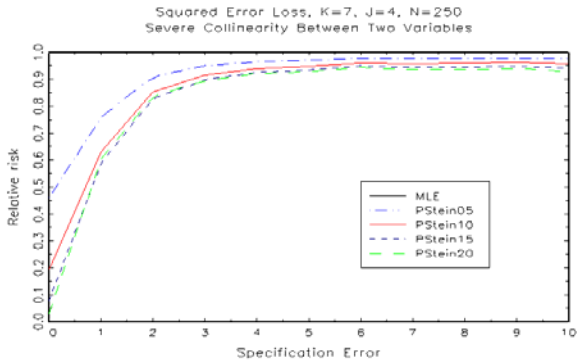
K = 4



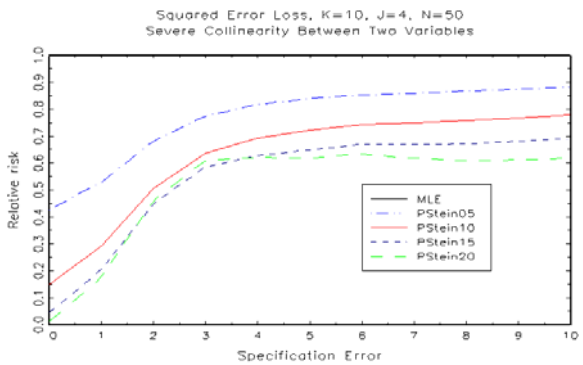


K = 7





K = 10



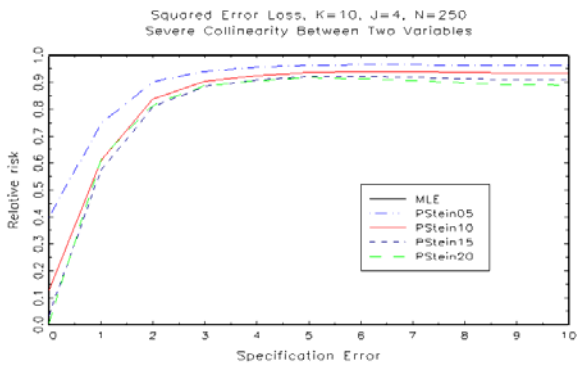
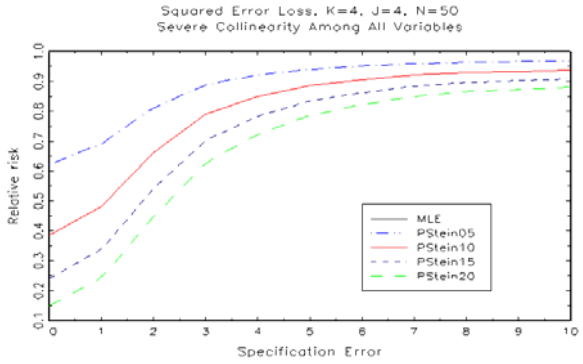
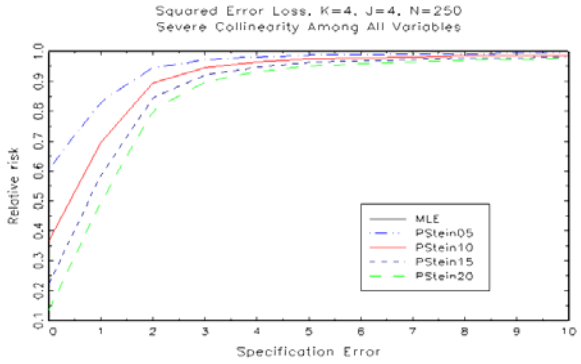


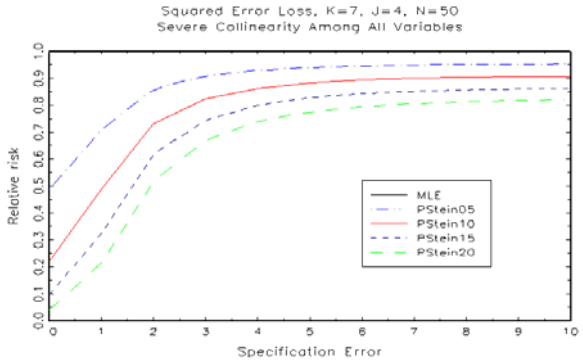
Figure 3.1: Squared Error Loss, Four Equally Likely Alternatives, Severe Collinearity Between Two Variables

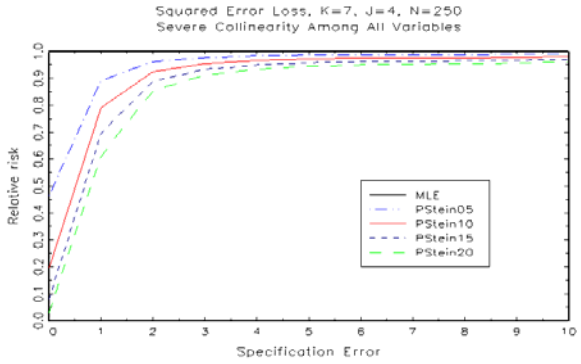
K = 4



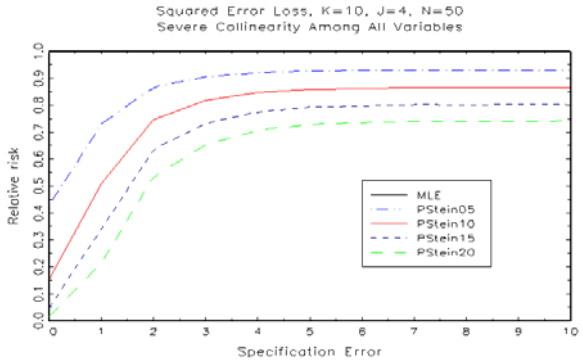


K = 7





K = 10



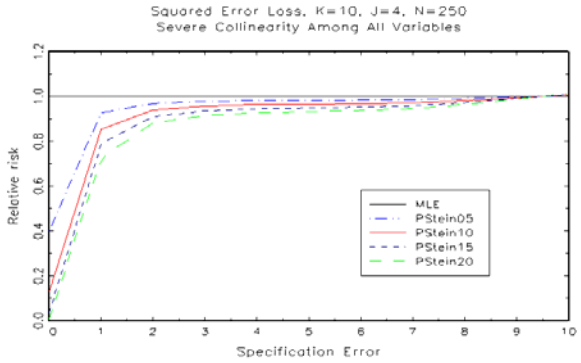


Figure 3.2: Squared Error Loss, Four Equally Likely Alternatives, Severe Collinearity Among All Variables

Compared to the relevant cases of low collinearity, the risk increase is 75% when severe collinearity is present between two variables, and more than 200% when severe collinearity is present among all variables. Although there is a significant percentage change in risk between the models, the values of the risk are relatively small. In all cases the risk decreases by more than 80% as the sample size increases which complies with asymptotic theory: as the sample size increases, both the bias and the variance of the estimators decrease. This is also confirmed by the estimated bias which is smaller when N=250.

Increasing the number of variables also increases the benefits of shrinkage. There is a larger risk improvement compared to the model with 4 variables and the risk differences are statistically significant over the entire parameter space, again with the exception of Pstein05 but only at w=1. The Stein rule estimators do not converge to the MLE for the parameter space that we consider. They will converge as we increase the length of the vector of true parameters β but we did not investigate the length at which it will occur. In terms of individual performance more shrinkage offers bigger risk improvement with the exception of the maximum shrinkage estimator Pstein20 which has the same risk as Pstein15 for K=7 and higher risk for K=10 for values of Estrella-R^2 smaller than .2. As we increase the size of the true parameters, the risk of Pstein20 decreases and remains the lowest among all estimators. We do not observe such behavior in the case of collinearity among all variables: there is a clear distinction between the Stein rule estimators and more shrinkage is better in terms of relative risk. In both cases the largest risk improvement over MLE is achieved in the model with 10 regressors. This was the result we obtained also in the orthonormal case. Stein rule estimators perform

better as we increase the number of restrictions, and more variables imply more restrictions since we shrink the parameters towards the null vector. Increasing the sample size shows similar performance in terms of relative risk as in the case of four variables when collinearity is present between two variables. In the case of collinearity among all variables, when K=10 the risk of the Stein estimators exceeds the risk of MLE for w=1, which corresponds to values of Estrella-R^2 of .9. The difference is small (the risk ratio is 1.01) and at this weight the number of samples excluded from the analysis is unacceptably high. The only exception is the minimum shrinkage estimator Pstein05, whose risk does not exceed the risk of MLE for all levels of specification error. In terms of magnitude, when N=50 we observed a significant jump in risk for the model where all variables are strongly correlated and K=10. The risk of the MLE reached its highest value of 93 at the largest degree of specification error in the modified model, compared to 5.8 if collinearity is present only between two variables. In the original model, where no samples are excluded, the magnitude of risk reaches five-digit numbers. When the model becomes more complex and severe collinearity is present, the small sample performance of all estimators shows that the obtained values for the coefficients are very far from the true values and the results are unreliable. Even if shrinkage improves estimation significantly, it is not good enough to justify estimating complex models with collinear data using only 50 observations. The models with N=250 show that the risk of all estimators remains relatively small even when K=10. Shrinkage offers significant risk improvement only for values of Estrella-R^2 less than .2, but the performance of all estimators improves.

For the models with low collinearity, the results are similar to the ones described above. When collinearity is present between two variables, the behavior of Pstein20 is more erratic than in the corresponding case of severe collinearity. Its risk becomes higher than the risk of the other shrinkage estimators for *w*=.3 when the estimators level out and start converging to the MLE. If *N*=250, the risk of Pstein20 even exceeds the risk of MLE for *w*=.1 or *w*=.2, but the risk differences are not statistically significant. When collinearity is present among all variables, comparable to the case of severe collinearity, there is a clear distinction between the Stein rule estimators and more shrinkage offers larger risk improvement. Overall, shrinkage offers significant risk improvement over a larger parameter space when collinearity is severe.

- Weighted Squared Error Loss

The effects of multicollinearity in choice models affect the maximum likelihood estimator through $\left[I\left(\hat{\beta}\right) \right]^{-1}$, which is the inverse of the information matrix, evaluated at MLE. The weighted squared error loss measures the distance of the estimator from the true parameter, weighted by the information matrix of the vector of parameters β. Since the information matrix is the inverse of the covariance matrix, the weighted loss will be inversely related to the variance of the parameters, i.e. more weight will be placed on parameters with smaller variance. In other words, the weighted loss penalizes more for errors in coefficients with smaller variability. Since shrinkage reduces the variability of the estimates, we expect the values of the weighted loss to be higher than the values of the squared loss, and the differences to be more pronounced for smaller degree of specification error. Table 3.7 shows the relative risk values in terms of weighted squared error loss, where severe collinearity is present between two variables.

Table 3.7: Weighted Squared Error Loss, Four Equally Likely Alternatives, Severe Collinearity Between Two Variables

	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	4.41	4.35	4.20	4.45	4.45	4.73	4.60	4.61	4.89	5.19	4.99
	PStein c=0.5	0.62	0.68	0.79	0.87	0.91	0.93	0.94	0.95	0.95	0.95	0.96
	PStein c=1	0.39	0.48	0.67	0.80	0.86	0.89	0.90	0.90	0.91	0.92	0.92
	PStein c=1.5	0.24	0.35	0.60	0.77	0.84	0.87	0.88	0.87	0.88	0.88	0.89
	PStein c=2	0.15	0.28	0.57	0.77	0.85	0.87	0.87	0.86	0.86	0.86	0.87
N=250	Risk MLE	4.05	3.94	4.09	4.17	4.13	4.12	4.25	4.02	4.19	4.25	4.23
	PStein c=0.5	0.61	0.83	0.95	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.37	0.74	0.92	0.96	0.97	0.98	0.98	0.98	0.98	0.98	0.98
	PStein c=1.5	0.22	0.71	0.93	0.97	0.97	0.97	0.97	0.98	0.98	0.98	0.97
	PStein c=2	0.14	0.71	0.96	0.98	0.97	0.97	0.97	0.97	0.97	0.97	0.97
						K=7						
N=50	Risk MLE	8.13	7.97	8.14	8.42	8.58	9.16	9.47	10.0	10.4	11.9	12.3
	PStein c=0.5	0.49	0.58	0.73	0.82	0.86	0.88	0.90	0.91	0.91	0.92	0.92
	PStein c=1	0.22	0.37	0.60	0.71	0.77	0.80	0.81	0.83	0.84	0.85	0.85
	PStein c=1.5	0.09	0.27	0.55	0.68	0.72	0.74	0.75	0.77	0.78	0.78	0.79
	PStein c=2	0.04	0.23	0.57	0.72	0.73	0.72	0.72	0.73	0.73	0.73	0.74
N=250	Risk MLE	7.16	7.05	7.04	7.19	7.22	7.22	7.15	7.34	7.43	7.43	7.45
	PStein c=0.5	0.46	0.79	0.92	0.95	0.97	0.97	0.98	0.98	0.98	0.98	0.98
	PStein c=1	0.20	0.71	0.89	0.93	0.94	0.95	0.96	0.96	0.96	0.97	0.96
	PStein c=1.5	0.08	0.74	0.90	0.92	0.93	0.94	0.95	0.94	0.95	0.95	0.95
	PStein c=2	0.03	0.82	0.96	0.94	0.93	0.93	0.94	0.93	0.93	0.94	0.94
						K=10						
N=50	Risk MLE	12.2	12.1	12.6	13.1	14.1	15.7	15.6	17.4	19.2	22.2	27.9
	PStein c=0.5	0.43	0.53	0.69	0.77	0.82	0.84	0.86	0.87	0.88	0.89	0.90
	PStein c=1	0.15	0.31	0.53	0.64	0.69	0.72	0.75	0.77	0.78	0.79	0.81
	PStein c=1.5	0.05	0.23	0.49	0.59	0.62	0.64	0.66	0.68	0.69	0.71	0.73
	PStein c=2	0.01	0.22	0.53	0.62	0.60	0.59	0.60	0.61	0.62	0.63	0.65
N=250	Risk MLE	10.2	10.3	10.7	10.4	10.4	10.6	10.9	10.8	11.1	11.1	11.5
	PStein c=0.5	0.40	0.77	0.91	0.94	0.96	0.96	0.97	0.97	0.97	0.97	0.97
	PStein c=1	0.13	0.68	0.86	0.90	0.92	0.94	0.94	0.94	0.94	0.95	0.95
	PStein c=1.5	0.04	0.72	0.85	0.89	0.91	0.92	0.92	0.92	0.92	0.93	0.93
	PStein c=2	0.01	0.84	0.89	0.89	0.90	0.91	0.90	0.90	0.90	0.91	0.91

Table 3.8 shows the relative risk values in terms of weighted squared error loss, where severe collinearity is present among all variables. In both cases all alternatives are equally likely.

Table 3.8: Weighted Squared Error Loss, Four Equally Likely Alternatives, Severe Collinearity Among All Variables

	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	4.41	4.34	4.25	4.56	4.53	4.78	4.93	5.15	5.17	5.29	5.49
	PStein c=0.5	0.62	0.73	0.87	0.92	0.94	0.95	0.96	0.96	0.96	0.96	0.96
	PStein c=1	0.39	0.57	0.80	0.87	0.89	0.91	0.92	0.92	0.93	0.93	0.93
	PStein c=1.5	0.24	0.47	0.77	0.84	0.86	0.88	0.89	0.89	0.90	0.90	0.90
	PStein c=2	0.15	0.41	0.78	0.84	0.85	0.86	0.88	0.87	0.87	0.87	0.88
N=250	Risk MLE	4.05	3.92	4.01	4.16	4.01	4.11	4.17	4.24	4.12	4.20	4.28
	PStein c=0.5	0.61	0.90	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.37	0.85	0.96	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
	PStein c=1.5	0.22	0.85	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98
	PStein c=2	0.14	0.88	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
						K=7						
N=50	Risk MLE	8.13	8.19	8.65	9.75	10.9	12.8	14.0	14.5	*14.1*	*12.6*	*10.8*
	PStein c=0.5	0.49	0.76	0.87	0.90	0.92	0.93	0.93	0.94	*0.94*	*0.94*	*0.94*
	PStein c=1	0.22	0.63	0.78	0.82	0.84	0.86	0.87	0.87	*0.88*	*0.89*	*0.89*
	PStein c=1.5	0.09	0.60	0.73	0.76	0.78	0.80	0.81	0.82	*0.83*	*0.83*	*0.84*
	PStein c=2	0.04	0.63	0.72	0.72	0.73	0.75	0.76	0.77	*0.78*	*0.79*	*0.80*
N=250	Risk MLE	7.16	7.11	7.15	7.31	7.52	7.47	7.76	7.71	8.11	7.61	*7.14*
	PStein c=0.5	0.46	0.92	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.99	*0.99*
	PStein c=1	0.20	0.89	0.94	0.96	0.96	0.97	0.97	0.97	0.97	0.97	*0.98*
	PStein c=1.5	0.08	0.89	0.93	0.94	0.95	0.95	0.95	0.96	0.95	0.96	*0.97*
	PStein c=2	0.03	0.93	0.93	0.93	0.93	0.94	0.94	0.94	0.94	0.95	*0.96*
						K=10						
N=50	Risk MLE	12.2	12.7	16.7	22.6	38.2	55.5	*57.9*	*63.5*	*44.6*	*39.6*	*31.9*
	PStein c=0.5	0.43	0.75	0.86	0.89	0.91	0.92	*0.92*	*0.92*	*0.92*	*0.92*	*0.92*
	PStein c=1	0.15	0.61	0.74	0.79	0.82	0.84	*0.84*	*0.85*	*0.85*	*0.85*	*0.85*
	PStein c=1.5	0.05	0.57	0.66	0.70	0.74	0.76	*0.77*	*0.78*	*0.78*	*0.78*	*0.78*
	PStein c=2	0.01	0.62	0.60	0.63	0.67	0.69	*0.70*	*0.71*	*0.71*	*0.71*	*0.72*
N=250	Risk MLE	10.2	10.3	10.8	11.1	11.6	11.7	*11.2*	*9.10*	*8.47*	*8.29*	*8.76*
	PStein c=0.5	0.40	0.94	0.97	0.97	0.98	0.98	*0.98*	*0.99*	*1.00*	*1.00*	*1.01*
	PStein c=1	0.13	0.90	0.94	0.95	0.95	0.96	*0.96*	*0.98*	*1.00*	*1.01*	*1.02*
	PStein c=1.5	0.04	0.89	0.92	0.93	0.93	0.94	*0.94*	*0.97*	*1.00*	*1.02*	*1.03*
	PStein c=2	0.01	0.90	0.90	0.91	0.91	0.92	*0.93*	*0.96*	*1.00*	*1.02*	*1.04*

In addition, because variability increases when the coefficients of the model increase in value, the risk increases but the errors will be weighted less compared to cases with smaller values of the coefficients. Therefore, we expect the values of weighted risk to be similar for each degree of hypothesis error. When collinearity is present among all

variables, we expect higher variance compared to a case of collinearity between two variables. With information matrix weighted error loss, we expect smaller weights in the first case, and larger weights in the second case, resulting in smaller differences in risk values compared to the squared error loss. The results confirm our expectations. The values of weighted error loss are larger than the values of the squared error loss in all cases, and the differences are more pronounced for small degrees of specification error, when we weigh more heavily the restricted estimates and their variability is lower. The differences in weighted error loss for different values of the true β are smaller compared to the differences in squared error loss. In addition, the values of the weighted error loss do not change much when collinearity is present among all variables, compared to collinearity between two variables. In both cases risk is smaller in large samples and increases as we add more variables to the model. In terms of relative risk Stein rule estimators dominate the MLE for the entire parameter space. When N=50 the shrinkage estimators do not converge to the MLE for the chosen length of the vector of coefficients. All risk differences are statistically significant, with the exception of Pstein05 a small percentage of the time. The risk improvement of Stein rule estimators over MLE is larger for models with more variables. The plots of the risk functions are similar to the case of squared error loss and are not presented. The results show that the shrinkage estimators approach faster the MLE when all variables are correlated. When K=4, this occurs at values of Estrella-R^2 close to .03, compared to Estrella-R^2 of .10 when collinearity is present only between two variables. We observe the same pattern also for models with K=7 and K=10. As we increase the sample size, the risk improvement over MLE is significant only over a small range of parameter space, and the risk of the shrinkage

estimators gets closer to the risk of MLE for values of Estrella-R^2 close to .10 for K=4, .4 for K=7, and .6 for K=10, when collinearity exists between two variables. When all variables are correlated, the results are comparable, but convergence generally occurs faster. In the latter case, when K=10 and N=250 the risk of the shrinkage estimators exceeds the risk of MLE when w=.8, although the highest risk ratio is 1.04 for Pstein20 and w=1, and this is in the range of values where higher than acceptable number of Monte Carlo samples is replaced and the results may not be representative of our experiment. In terms of relative performance of the Stein rule estimators, more shrinkage leads to bigger risk improvement in small samples, with the exception of Pstein20 whose risk does not increase monotonically and intersects the risk of other estimators usually when the weight is as low as .2, after which decreases again and becomes lower than the risk of the rest of the estimators. In large samples the results are similar, but the differences between the estimators are less distinct.

Unlike the case of squared error loss, there are no significant differences between the relevant models with severe and low correlation. We expect variability to be higher if correlation is severe than if correlation is low, but there will be different weights in the two cases, which explains the similarity in the results.

- Out-Of-Sample Prediction Loss

The results for prediction out-of-sample are presented in Tables 3.9 and 3.10. For each Monte Carlo experiment we generate a sample of 100 observations which are not used in the estimation of the model parameters. The data generating process is the same as the one used in the original sample.

Table 3.9: Mean Squared Error of Prediction Out of Sample, Four Equally Likely Alternatives, Severe Collinearity Between Two Variables[16]

	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	2.08	2.04	1.97	2.06	2.04	2.09	1.97	1.93	2.00	2.01	1.95
	PStein c=0.5	0.63	0.68	0.81	0.90	0.94	0.97	0.98	0.98	0.98	0.98	0.99
	PStein c=1	0.39	0.49	0.69	0.84	0.92	0.95	0.97	0.97	0.97	0.97	0.98
	PStein c=1.5	0.24	0.37	0.64	0.83	0.94	0.97	0.98	0.97	0.97	0.97	0.98
	PStein c=2	0.15	0.29	0.62	0.86	0.98	1.01	1.01	0.99	0.99	0.98	0.98
N=250	Risk MLE	0.40	0.38	0.39	0.40	0.39	0.39	0.41	0.39	0.42	0.42	0.43
	PStein c=0.5	0.61	0.83	0.95	0.98	0.99	0.99	0.99	0.99	1.00	1.00	1.00
	PStein c=1	0.37	0.74	0.92	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
	PStein c=1.5	0.22	0.70	0.92	0.97	0.98	0.98	0.98	0.99	0.99	0.99	0.99
	PStein c=2	0.14	0.69	0.95	0.98	0.98	0.98	0.98	0.99	0.99	0.99	0.99
						K=7						
N=50	Risk MLE	3.92	3.78	3.81	3.85	3.84	3.90	3.94	4.02	4.03	4.20	4.25
	PStein c=0.5	0.50	0.60	0.76	0.86	0.91	0.94	0.95	0.96	0.96	0.97	0.97
	PStein c=1	0.23	0.39	0.64	0.79	0.86	0.90	0.91	0.93	0.93	0.94	0.94
	PStein c=1.5	0.10	0.29	0.61	0.79	0.85	0.88	0.90	0.91	0.91	0.92	0.92
	PStein c=2	0.04	0.25	0.64	0.86	0.90	0.91	0.90	0.91	0.91	0.90	0.91
N=250	Risk MLE	0.70	0.69	0.70	0.72	0.72	0.73	0.73	0.75	0.76	0.79	0.78
	PStein c=0.5	0.47	0.81	0.94	0.97	0.98	0.98	0.99	0.99	0.99	0.99	0.99
	PStein c=1	0.20	0.75	0.92	0.95	0.97	0.97	0.98	0.98	0.98	0.99	0.98
	PStein c=1.5	0.08	0.80	0.95	0.96	0.96	0.97	0.98	0.97	0.98	0.98	0.98
	PStein c=2	0.03	0.91	1.03	0.98	0.97	0.97	0.97	0.97	0.97	0.98	0.97
						K=10						
N=50	Risk MLE	5.71	5.64	5.71	5.66	5.70	5.74	5.55	5.65	5.75	5.90	6.06
	PStein c=0.5	0.45	0.57	0.76	0.86	0.91	0.93	0.94	0.95	0.96	0.96	0.97
	PStein c=1	0.16	0.37	0.65	0.80	0.85	0.88	0.90	0.92	0.93	0.93	0.94
	PStein c=1.5	0.05	0.30	0.68	0.84	0.86	0.87	0.89	0.89	0.90	0.91	0.91
	PStein c=2	0.02	0.29	0.78	0.98	0.94	0.90	0.90	0.89	0.89	0.89	0.89
N=250	Risk MLE	0.99	1.02	1.10	1.10	1.12	1.18	1.21	1.21	1.25	1.26	1.31
	PStein c=0.5	0.40	0.78	0.92	0.95	0.97	0.98	0.98	0.99	0.99	0.99	0.99
	PStein c=1	0.13	0.69	0.87	0.92	0.95	0.96	0.97	0.97	0.98	0.98	0.98
	PStein c=1.5	0.04	0.72	0.86	0.91	0.94	0.95	0.96	0.96	0.97	0.97	0.97
	PStein c=2	0.01	0.84	0.90	0.92	0.94	0.95	0.95	0.96	0.96	0.96	0.97

[16] The risk of MLE is multiplied by 100 in all tables for out-of-sample prediction.

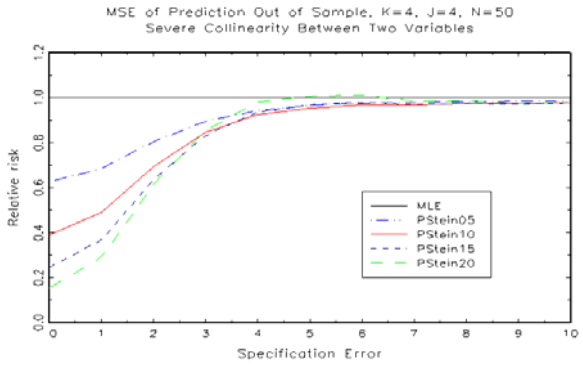
Table 3.10: Mean Squared Error of Prediction Out of Sample, Four Equally Likely Alternatives, Severe Collinearity Among All Variables

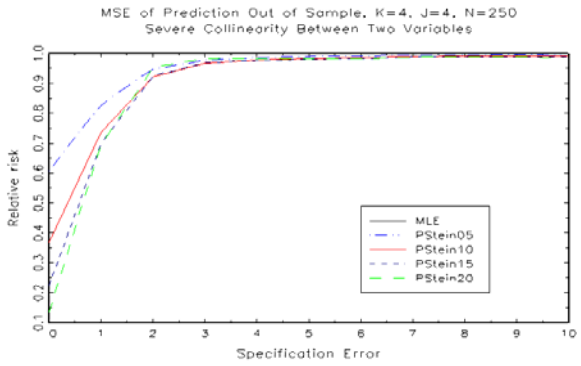
	W	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
						K=4						
N=50	Risk MLE	2.08	2.04	2.02	2.16	2.13	2.20	2.21	2.17	2.18	2.19	2.17
	PStein c=0.5	0.63	0.74	0.89	0.94	0.96	0.97	0.98	0.98	0.99	0.99	0.99
	PStein c=1	0.39	0.58	0.83	0.91	0.94	0.96	0.97	0.97	0.97	0.97	0.98
	PStein c=1.5	0.24	0.48	0.81	0.90	0.93	0.95	0.97	0.96	0.97	0.96	0.97
	PStein c=2	0.15	0.42	0.83	0.91	0.94	0.95	0.97	0.96	0.96	0.96	0.96
N=250	Risk MLE	0.40	0.38	0.39	0.41	0.40	0.41	0.43	0.44	0.44	0.45	0.47
	PStein c=0.5	0.61	0.90	0.97	0.99	0.99	0.99	0.99	1.00	1.00	1.00	1.00
	PStein c=1	0.37	0.86	0.96	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	PStein c=1.5	0.22	0.86	0.97	0.98	0.99	0.98	0.99	0.99	0.99	0.99	0.99
	PStein c=2	0.14	0.90	0.99	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99
						K=7						
N=50	Risk MLE	3.92	4.03	4.23	4.34	4.50	4.76	4.86	5.05	*5.26*	*5.34*	*5.24*
	PStein c=0.5	0.50	0.79	0.92	0.96	0.97	0.98	0.98	0.98	*0.99*	*0.99*	*0.99*
	PStein c=1	0.23	0.68	0.88	0.93	0.94	0.96	0.96	0.97	*0.97*	*0.97*	*0.98*
	PStein c=1.5	0.10	0.67	0.87	0.91	0.92	0.94	0.95	0.95	*0.96*	*0.96*	*0.97*
	PStein c=2	0.04	0.71	0.90	0.90	0.91	0.93	0.93	0.94	*0.95*	*0.95*	*0.96*
N=250	Risk MLE	0.70	0.73	0.75	0.76	0.78	0.78	0.80	0.81	0.85	0.85	*0.87*
	PStein c=0.5	0.47	0.93	0.98	0.99	0.99	0.99	0.99	0.99	0.99	1.00	*1.00*
	PStein c=1	0.20	0.89	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99	*0.99*
	PStein c=1.5	0.08	0.89	0.96	0.98	0.98	0.99	0.99	0.99	0.99	0.99	*0.99*
	PStein c=2	0.03	0.93	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98	*0.99*
						K=10						
N=50	Risk MLE	5.71	5.58	5.91	6.22	7.03	7.65	*8.01*	*8.59*	*8.39*	*8.38*	*8.32*
	PStein c=0.5	0.45	0.82	0.94	0.96	0.97	0.98	*0.98*	*0.98*	*0.99*	*0.99*	*0.99*
	PStein c=1	0.16	0.74	0.89	0.93	0.94	0.96	*0.96*	*0.97*	*0.97*	*0.97*	*0.98*
	PStein c=1.5	0.05	0.76	0.87	0.90	0.92	0.94	*0.94*	*0.95*	*0.96*	*0.96*	*0.96*
	PStein c=2	0.02	0.88	0.88	0.88	0.90	0.92	*0.93*	*0.94*	*0.94*	*0.95*	*0.95*
N=250	Risk MLE	0.99	1.07	1.15	1.19	1.26	1.27	*1.29*	*1.19*	*1.21*	*1.25*	*1.35*
	PStein c=0.5	0.40	0.95	0.98	0.99	0.99	0.99	*0.99*	*1.00*	*1.00*	*1.01*	*1.01*
	PStein c=1	0.13	0.93	0.97	0.98	0.98	0.99	*0.99*	*1.00*	*1.00*	*1.01*	*1.02*
	PStein c=1.5	0.04	0.93	0.96	0.97	0.98	0.98	*0.98*	*0.99*	*1.01*	*1.02*	*1.03*
	PStein c=2	0.01	0.95	0.96	0.96	0.97	0.97	*0.98*	*0.99*	*1.01*	*1.03*	*1.04*

The true probabilities are computed as $P_{ij} = \frac{\exp(z'_{Oij}\beta)}{\sum_{j=1}^{J}\exp(z'_{Oij}\beta)}$, and the predicted probabilities are computed as $\hat{P}_{ij} = \frac{\exp(z'_{Oij}\hat{\beta})}{\sum_{j=1}^{J}\exp(z'_{Oij}\hat{\beta})}$, where $\hat{\beta}$ takes the values obtained by each of the nine estimators using the original sample, and Z_O is the matrix of observations generated for out-of-sample prediction. We use two different loss functions to evaluate out-of-sample performance, MSE_P and MSE_Y, described in detail in the orthonormal case. The results in the tables are obtained using MSE_P. Since the values of risk are very small, the risk of the MLE is multiplied by 100 in the output tables to be able to make comparison between models. The conventional wisdom about the effects of multicollinearity on prediction in linear regression tells us that the risk of prediction should not be affected by correlations among the regressors, as long as the values of the explanatory variables Z_O obey the same pattern of multicollinearity as the sample data *Z*. Therefore, we expect the performance of MLE and the shrinkage estimators as out-of-sample predictors not to be affected by the presence of collinearity. Consequently, we expect to obtain similar estimates of prediction risk in cases of different patterns or degrees of collinearity. The results show that the values of MSE_P are similar in the case of severe collinearity between two and among all variables of the model, and also similar to the results in the orthonormal case. In all cases the risk estimates do not differ much as we change the degree of specification error, and increase as we increase the number of explanatory variables. The risk increases more when *N*=50 and reaches its highest values for *K*=10 and collinearity among all variables, where the model becomes too complex to

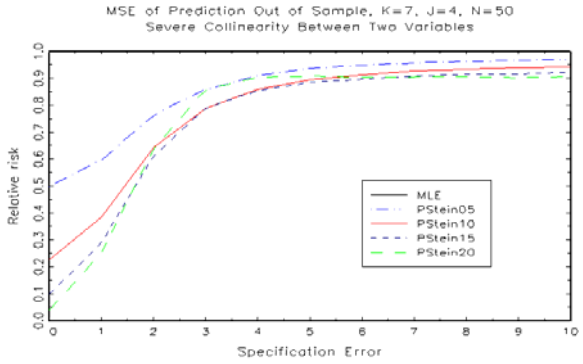
be estimated by such a small number of observations. The results for relative risk show that shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. For *N*=50 Stein rule estimators show larger risk improvement when collinearity is present between two variables. The risk of the shrinkage estimators starts converging to the risk of MLE towards the mid-range of specification error, or for values of Estrella-R^2 equal to .12 for *K*=4, .27 for *K*=7, and .35 for *K*=10. The risk of Pstein20 exceeds the risk of the other shrinkage estimators close to convergence and decreases again afterwards. The spike in the risk function is the highest for *K*=10, as shown in Figure 3.3. In the case of collinearity among all variables, convergence to MLE occurs faster and the risk of Pstein20 is closer to the risk of the other shrinkage estimators, although again no estimator is dominant over the entire parameter space. For *N*=250 shrinkage offers risk improvement only for small degree of specification error, and again performance of Stein rule estimators is better for collinearity between two variables. If we use MSE_Y, we replace the true probability with the actual choice of an alternative, where $y_{ij} = 1$ for the alternative with the highest utility, and zero otherwise. In this case we observe a decrease in the risk of all estimators as the signal-to-noise ratio increases. The risk differences are small, ranging from .77 for *w*=0 to .54 for *w*=1 in the model with *K*=4 and *N*=50, and the values for the other models differ only slightly from these results. When we shrink the model parameters to zero we imply that all alternatives are equally likely, therefore $P_{ij} = 1/J$. As we move away from this restriction, we introduce more information to the model through the estimated coefficients and the explanatory variables, and we allow for a larger range of predicted probabilities.

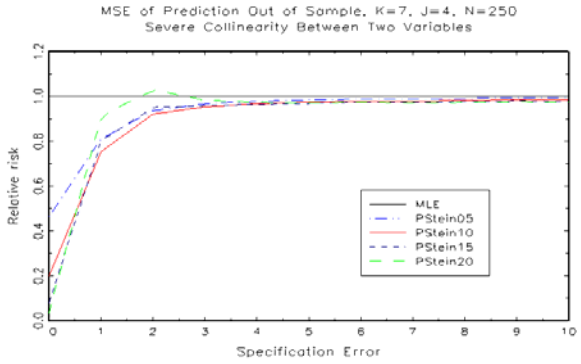
K = 4



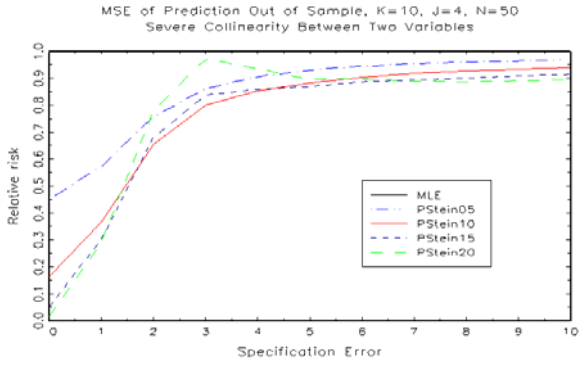


K = 7





K = 10



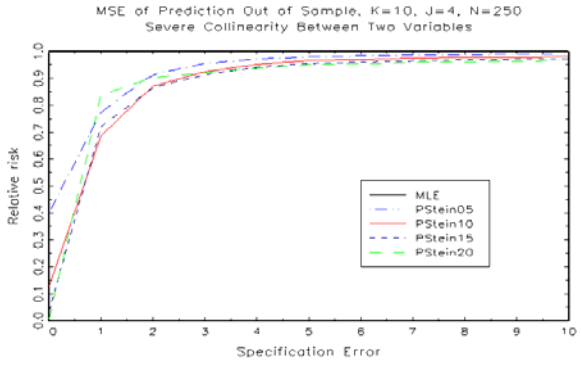


Figure 3.3: MSE of Prediction Out-Of-Sample, Four Equally Likely Alternatives, Severe Collinearity Between Two Variables

Since in the data generation process we set $y_{ij} = 1$ for the alternative associated with the highest utility, and both the utilities and the predicted probabilities depend on the values of the explanatory variables, it becomes easier to predict an outcome as the signal-to-noise ratio increases. The hit rate confirms this conclusion. In general, we expect the hit rate to behave similarly to MSE_Y, because both measures depend on the ability to predict $y_{ij} = 1$ when alternative j is chosen. There are no significant differences in estimator performance for the relevant models of severe and low collinearity, because the presence and degree of collinearity generally does not affect out-of-sample prediction, as long as the sample data and the non-sample data follow the same pattern of collinearity.

3.4.1.2 Four Alternatives With One Dominant

The model with one dominant alternative represents cases in which one of the alternatives is chosen by the majority of individuals. Since the probability of choosing an alternative depends on the values of the explanatory variables, we created a dominant alternative by changing the mean of the explanatory variables associated with this alternative. It was achieved in the Monte Carlo experiment by adding a constant to the explanatory variables of the dominant alternative. In most cases the constant was equal to 0.2 and generated an alternative which was selected by 80% of the individuals when the values of the true β equal to one. The value of the constant in the orthonormal case was 0.1. In our experiment this implies that when collinearity is present the explanatory variables have smaller influence over a choice of an alternative, compared to the orthonormal case. For smaller degrees of hypothesis error the share of the dominant alternative is smaller, and all alternatives are equally likely when specification error is zero and the true β is determined only by the restricted model.

- Goodness of Fit

The goodness of fit measures improve compared to the model with four equally likely alternatives, and the improvement is larger as we increase the signal-to-noise ratio, which corresponds to larger share of the dominant alternative. This is because as the probability of choosing an alternative increases, the value of the log-likelihood increases as well. As we increase the number of variables, the values of Estrella-R^2 get close to one. The chances of a probability being equal to one or to zero increase, which could make the model a perfect predictor and may cause the estimation to break down, as explained in the orthonormal case. To avoid this problem we excluded from the Monte Carlo experiment samples with values of Estrella-R^2 greater than 0.98, and the number of excluded samples for each model is shown in Table 3.13. The model with 10 variables and 250 observations was estimated without the exclusion of Monte Carlo samples because of the long computation time.

- Squared Error Loss

The results for relative risk are very similar to the model with equally likely alternatives and show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error. Shrinkage leads to risk improvement also in this case due to the following reasons. For small degrees of hypothesis error the parameters of the model are close to zero and all alternatives have similar shares. In particular, each alternative get chosen 25% of the time when the true β is a vector of zeros. In these cases we place more weight on the restricted estimator, which significantly reduces estimation risk. As the model coefficients increase in value the differences between the shares of each alternative increase as well. In these cases the

value of the likelihood ratio test statistic is large and we place more weight on the unrestricted estimator, therefore we are not imposing the restriction of equally likely alternatives. Generally, for severe collinearity between two variables, when the share of the dominant alternative exceeds 40% for N=50 and 30% for N=250, the likelihood ratio test statistic is statistically significant and we can reject the hypothesis of a null vector of coefficients. In the case of severe collinearity among all variables the corresponding numbers are close to 35% for N=50 and close to .28 for N=250. The plots of the risk functions are very similar to the cases with equally likely alternatives, and therefore are not presented. The only difference we observed was for the model with K=7, N=50, and collinearity between two variables. In the case of equally likely alternatives there was no dominant estimator, while if one alternative is dominant there is a clear distinction between the estimators and more shrinkage means larger risk improvement over the parameter space under consideration. In addition, where K=10, N=250 and collinearity is present between two variables, when w=.9, corresponding to adjusted values of Estrella-R^2 equal to .98, the risk of the shrinkage estimators exceeds the risk of MLE. However, at this weight the results cannot be considered representative, because the number of excluded sample is 4,582, followed by 663,201 for w=1. As in the case with equally likely alternatives, the results show that the risk improvement is larger for N=50 and for models with more explanatory variables. The risk differences for N=50 are statistically significant with the exception for Pstein05 and Pstein20 a percentage of the time. When the number of variables increases, the risk differences are statistically significant also for N=250, again with exceptions similar to the previous case. In terms of magnitude, the risk of all estimators is larger compared to the case of equally likely alternatives, and

when collinearity is present among all variables, compared to the case of two variables. In both cases of collinearity the risk decreases as the sample size increases. Increasing the number of variables shows that the models become increasingly difficult to estimate in small samples. The risk increases almost 20-fold when K=10 in the adjusted model, and much more in the original model, where no samples are excluded. When the model becomes more complex and severe collinearity is present, the small sample performance of all estimators shows that the obtained values for the coefficients are very far from the true values and the results are unreliable. The models with N=250 show that the risk of all estimators remains relatively small even when K=10, although the number of excluded samples increases substantially. We did not estimate an adjusted model for K=10 and N=250 because of the long computation time. The program was interrupted after more than 36 hours of execution. The relative risk does not change as a result of replacing Monte Carlo samples, therefore we can be confident when we compare these results to the rest of the cases. Comparing out results to the case of low collinearity does not add additional information. The relative risk functions behave in a manner similar to the cases of severe collinearity, and generally the values of loss are lower with low collinearity, which is a result we expect.

- Weighted Squared Error Loss

The results for weighted error loss are similar to the case with four equally likely alternatives and show that the Stein rule estimators dominate the MLE for the entire parameter space, and the risk differences improve in statistical significance as the number of variables increases. The improvement is more significant when collinearity is present between two variables. In all cases there is no dominant shrinkage estimator. The risk

functions are very similar to the case of equally likely alternatives, therefore the graphs are not included in the appendix. In terms of magnitude, the risk values do not differ much with the degree of specification error, or with the type or degree of collinearity. The risk in all cases increases when we add more explanatory variables to the models.

- Out-Of-Sample Prediction Loss

The results for out-of-sample prediction show that there are no significant differences from the case of equally likely alternatives, and between the two types of collinearity. Overall the risk values are very small which shows that all estimators perform very well as predictors out-of-sample. The graphs of the risk functions are very similar to the ones obtained in the case of equally likely alternatives, and are not included. The only difference we found was for the model with K=10, N=50, and severe collinearity between two variables, where the spike in the risk of Pstein20 was smaller than in the case of equally likely alternatives. The hit rate shows very good predictive ability out-of-sample, and much better performance compared to the model with equally likely alternatives. In the case with 4 variables and 50 observations, we start observing differences in the hit rate when the share of the dominant alternative exceeds 30%, and they increase as the differences in shares become more apparent. The hit rate for w=1 reaches 87%. These results are expected, because it is easier to predict correctly an outcome where one of the alternatives is dominant. Generally all estimators over-predict the dominant alternative, which is the results we obtained also in the orthonormal case. The values of the hit rate for all estimators are the same and there are no significant differences as we increase the sample size, the number of variables, or if we change the form or degree of collinearity in the model.

3.4.1.3 Four Alternatives With Two Dominant

The model with two dominant alternatives represents cases in which two of the alternatives are chosen by the majority of individuals, and the two alternatives have similar shares. In the next chapter we estimate a model which represents such a situation. Using the same reasoning about the choice of an alternative as in the previous case, we create two dominant alternatives by adding a constant (equal to 0.15 in most cases) to the explanatory variables of the dominant alternatives. The common share of the two dominant alternatives is about 90% when the values of the true β equal to one. Like in the previous case, for smaller degrees of hypothesis error the shares of the dominant alternatives are smaller, and all alternatives are equally likely when specification error is zero and the true β is determined only by the restricted model.

- Goodness of Fit

The goodness of fit measures are generally lower than in the model with one dominant alternative, because the chances of a probability being equal to one is much smaller in this case. The number of excluded samples is generally smaller compared to the model with one dominant alternative, and also smaller for models with collinearity only between two variables.

- Squared Error Loss

The results for relative risk are very similar to the models with one dominant or equally likely alternatives and show that under squared error loss Stein rule estimators dominate the MLE for each sample size and each level of specification error, for the same reasons discussed in the case of one dominant alternative. At low degree of specification error we place more weight on the restricted estimator, but the actual shares are very

similar, and identical when $\beta = 0$. As the differences in shares for each alternative increase, we weigh heavily the unrestricted model and do not impose the restriction of equally likely alternatives. Generally, for severe collinearity between two variables, when the common share of the two dominant alternatives exceeds 66% for *N*=50 and 55% for *N*=250, the likelihood ratio test statistic becomes statistically significant and we can reject the hypothesis of a null vector of coefficients. In the case of severe collinearity among all variables the corresponding numbers are comparable but smaller. The plots of the risk functions are very similar to the cases with equally likely alternatives, and therefore are not included in the text. The only difference we observed was for the model with *K*=7, *N*=50, and collinearity between two variables. In the case of equally likely alternatives there was no dominant estimator, while if one alternative is dominant there is a clear distinction between the estimators and more shrinkage means larger risk improvement over the parameter space under consideration. This is the result we obtained also in the case of one dominant alternative. The risk differences for *N*=50 are statistically significant with the exception for Pstein05 and Pstein20 a percentage of the time. When the number of variables increases, the risk differences are statistically significant also for *N*=250, again with some similar exceptions. In terms of magnitude, the risk of all estimators is very similar to the risk when one alternative is dominant. In both cases of collinearity the risk decreases as the sample size increases. Increasing the number of variables shows that the models become increasingly difficult to estimate in small samples, and again we observe the highest risk values for *K*=10 and *N*=50. The models with *N*=250 show that the risk of all estimators remains relatively small even when *K*=10. We did not estimate an adjusted model for *K*=10 and *N*=250 because of the

long computation time. Comparing out results to the case of low collinearity does not add any important additional information.

- Weighted Squared Error Loss

As in the previous two cases Stein rule estimators dominate the MLE for the entire parameter space, and the risk improvement is larger in small samples and for models with more variables. More shrinkage leads to larger risk improvement, but in most cases the risk of Pstein20 exceeds the risk of the other shrinkage estimators, usually when they begin converging to the MLE in large samples, or level out in small samples. The risk functions do not differ from the model with equally likely alternatives, therefore the plots are not presented. The values of the weighted risk are very similar to the values when one alternative is dominant. The only more significant difference we found in the model of collinearity between two variables. When K=10 and N=50, the risk of MLE for the model with one dominant alternative increases faster than in the case of two dominant alternatives. In addition, much more Monte Carlo samples were excluded in the case of one dominant alternative, which makes the results for the last five weights not representative of our experiment.

- Out-Of-Sample Prediction Loss

The results for relative risk confirm out findings in the previous two cases: shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space. The plots of the relative risk are similar to the case of equally likely alternatives and are not presented. The only difference is again in the case of K=10, N=50, and collinearity between two variables, where the spike in Pstein20 is smaller than when the four alternatives are equally likely. In terms of magnitude, the risk

values are higher here than in the previous two cases when severe collinearity is present between two variables. When all variables are severely correlated, the risk differences between the models get smaller. The hit rate again shows very good predictive ability, although its values are lower than in the case of one dominant alternative. Again the values of the hit rate for all estimators are the same and there are no significant differences as we increase the sample size or the number of variables in the model. Overall, also in this case, all estimators predict very well out-of-sample, but the Stein estimators outperform the MLE for small to moderate degrees of hypothesis error.

3.4.2 Collinearity Between Alternative-Specific Variables

Collinearity between alternatives-specific variables was generated with the use of auxiliary regressions of the form presented in equation (3.3), and a transformation of the variables shown in equation (3.8). We analyzed models with 4, 7 and 10 alternatives where one variable was correlated between two, three, and four alternatives, respectively. In a marketing context we were thinking about prices, which may be correlated between different brands.

We studied the correlations between variables when the data matrix is arranged in two different ways. First, the matrix (call it matrix A) has N rows and $J*K$ columns, i.e. there is a column for each alternative-specific variable. Second, a matrix B is in "stacked" form, i.e. it has dimensions $(N*J)$xK so there are J rows for each observation, where J is the number of alternatives. When we generate collinearity between alternative-specific variables, the desired degree of correlation appears in Matrix A, but does not appear in Matrix B. When the data are in stacked form the alternative-specific effects are lost and only correlations between variables matter. All software packages

estimate conditional model in stacked form, and therefore even if severe collinearity is present between alternative-specific variables, it will have no effect on the estimation results. In conditional logit the effect of a variable is assumed constant across alternatives and the model estimates a single coefficient associated with that variable.

- Goodness of Fit

Table 3.11 shows the goodness of fit measures where severe collinearity is present between alternative-specific variables. Table 3.12 shows the corresponding values in the case of low collinearity. All models were estimated in the original form, i.e. no samples with values of Estrella-R^2 higher than .98 were excluded from the simulation. Therefore, we observe goodness of fit values approaching one, especially in models with one or more dominant alternatives, which is an indication that we may observe the perfect classifier problem, especially when the sample size is small and the number of alternative increases. Overall, there is no difference between the models with severe collinearity and the ones with low collinearity.

- Squared Error Loss

Because of the large volume of presented information, and the similarity of the results, the numerical values when collinearity is present among alternative specific variables are not included but can be provided upon request. The results for absolute and relative risk are very similar to the results obtained in the orthonormal case. Due to the reasons explained above, since collinearity among alternatives does not affect the estimation results in conditional logit, this is an expected outcome. In terms of relative risk, shrinkage estimators dominate the maximum likelihood estimator over the entire parameter space.

Table 3.11: Goodness of Fit Measures, Severe Collinearity Among Alternatives[17]

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.05	0.04	0.17	0.30	0.43	0.53	0.62	0.69	0.75	0.79
			BIC	3.01	2.97	2.88	2.74	2.58	2.40	2.24	2.08	1.95	1.81	1.70
			LR	4.02	5.71	10.32	17.12	25.27	34.04	42.28	50.06	56.79	63.76	69.52
4	4	250	Estrella	-0.02	0.01	0.09	0.21	0.33	0.45	0.55	0.64	0.71	0.76	0.81
			BIC	2.84	2.82	2.73	2.61	2.46	2.29	2.14	1.98	1.84	1.71	1.58
			LR	3.99	11.21	32.09	63.45	101.34	141.59	180.93	220.22	256.29	288.74	319.01
7	4	50	Estrella	-0.08	-0.05	0.06	0.20	0.36	0.50	0.63	0.72	0.80	0.85	0.89
			BIC	4.13	4.09	3.98	3.81	3.60	3.38	3.14	2.91	2.69	2.48	2.31
			LR	3.99	5.87	11.28	19.52	30.07	41.11	53.34	64.59	75.66	86.14	94.94
7	4	250	Estrella	-0.02	0.02	0.11	0.24	0.38	0.52	0.63	0.72	0.79	0.84	0.88
			BIC	3.96	3.93	3.83	3.68	3.49	3.28	3.07	2.86	2.67	2.48	2.31
			LR	4.07	12.30	36.59	74.17	121.75	174.77	227.57	279.96	328.64	375.39	418.15
10	4	50	Estrella	-0.08	-0.05	0.04	0.17	0.32	0.45	0.58	0.68	0.76	0.82	0.87
			BIC	4.84	4.81	4.71	4.57	4.38	4.18	3.94	3.72	3.49	3.27	3.08
			LR	4.07	5.65	10.54	17.30	26.98	36.84	48.78	59.92	71.50	82.38	91.97
10	4	250	Estrella	-0.02	0.02	0.12	0.25	0.40	0.54	0.66	0.76	0.83	0.88	0.91
			BIC	4.68	4.64	4.54	4.38	4.17	3.94	3.68	3.44	3.20	2.98	2.77
			LR	4.04	12.91	38.65	79.26	130.62	189.55	252.20	312.74	374.17	428.67	479.87
4	4	50	Estrella	-0.08	-0.04	0.07	0.25	0.46	0.65	0.80	0.89	0.94	0.97	0.99
One dominant			BIC	3.01	2.97	2.84	2.64	2.36	2.02	1.68	1.37	1.09	0.84	0.64
			LR	4.02	5.93	12.06	22.14	36.46	53.37	70.04	86.02	99.84	112.36	122.34
4	4	250	Estrella	-0.02	0.02	0.12	0.28	0.47	0.65	0.79	0.89	0.94	0.98	0.99
One dominant			BIC	2.84	2.81	2.71	2.52	2.26	1.96	1.64	1.31	1.03	0.77	0.55
			LR	4.00	12.12	38.38	84.76	149.89	225.35	305.91	386.49	458.98	522.44	577.16
7	4	50	Estrella	-0.08	-0.04	0.08	0.26	0.48	0.67	0.83	0.92	0.97	0.99	1.00
One dominant			BIC	4.12	4.08	3.96	3.75	3.43	3.04	2.58	2.13	1.72	1.35	1.04
			LR	4.00	6.00	12.24	22.88	38.97	58.47	81.40	103.82	124.26	142.87	158.23
7	4	250	Estrella	-0.02	0.02	0.12	0.29	0.48	0.67	0.82	0.92	0.97	0.99	1.00
One dominant			BIC	3.96	3.93	3.82	3.62	3.34	2.97	2.54	2.08	1.63	1.23	0.90
			LR	4.06	12.80	40.75	89.12	159.62	251.87	359.25	474.62	587.31	687.91	770.49
10	4	50	Estrella	-0.08	-0.05	0.06	0.20	0.40	0.60	0.77	0.89	0.96	0.99	1.00
One dominant			BIC	4.84	4.80	4.70	4.53	4.25	3.90	3.46	2.95	2.41	1.87	1.39
			LR	4.07	5.68	11.12	19.56	33.45	50.78	73.08	98.28	125.27	152.23	176.38
10	4	250	Estrella	-0.02	0.02	0.12	0.28	0.47	0.66	0.81	0.91	0.97	0.99	1.00
One dominant			BIC	4.68	4.64	4.53	4.34	4.07	3.70	3.25	2.75	2.22	1.71	1.27
			LR	3.94	13.07	40.77	88.12	156.78	248.88	360.64	486.22	618.32	744.84	856.71
4	4	50	Estrella	-0.08	-0.05	0.07	0.22	0.39	0.55	0.67	0.75	0.81	0.86	0.90
Two dominant			BIC	3.01	2.97	2.85	2.68	2.46	2.21	2.00	1.81	1.64	1.49	1.35
			LR	4.03	5.87	11.57	20.30	31.52	43.57	54.49	63.98	72.33	79.93	86.74
4	4	250	Estrella	-0.02	0.02	0.11	0.25	0.41	0.56	0.68	0.77	0.84	0.89	0.92
Two dominant			BIC	2.84	2.81	2.72	2.56	2.35	2.12	1.89	1.67	1.47	1.30	1.15
			LR	4.01	11.94	36.28	75.93	127.06	184.75	242.31	296.87	346.52	391.04	428.75
7	4	50	Estrella	-0.08	-0.04	0.07	0.23	0.42	0.58	0.72	0.82	0.89	0.93	0.95
Two dominant			BIC	4.13	4.08	3.97	3.78	3.52	3.24	2.92	2.62	2.34	2.07	1.87
			LR	3.99	5.99	11.79	21.31	34.09	47.99	64.28	79.23	93.34	106.82	116.83
7	4	250	Estrella	-0.02	0.02	0.12	0.26	0.43	0.58	0.71	0.80	0.87	0.92	0.95
Two dominant			BIC	3.96	3.93	3.82	3.65	3.43	3.16	2.89	2.61	2.34	2.09	1.87
			LR	3.98	12.74	39.45	82.26	138.52	204.43	272.69	342.74	409.48	471.41	527.13
10	4	50	Estrella	-0.08	-0.05	0.05	0.20	0.38	0.55	0.70	0.81	0.89	0.93	0.96
Two dominant			BIC	4.84	4.80	4.70	4.53	4.28	4.00	3.66	3.31	2.97	2.65	2.38
			LR	4.05	5.74	11.07	19.36	31.74	45.71	62.69	80.30	97.58	113.50	127.07
10	4	250	Estrella	-0.02	0.02	0.12	0.27	0.44	0.60	0.73	0.83	0.89	0.93	0.96
Two dominant			BIC	4.68	4.64	4.53	4.36	4.11	3.83	3.51	3.19	2.87	2.59	2.34
			LR	3.95	13.17	40.15	84.49	145.09	216.91	296.88	376.54	454.68	525.37	588.70

[17] The models are estimated without a check for values of Estrella-R^2.

Table 3.12: Goodness of Fit Measures, Low Collinearity Among Alternatives[18]

J	K	N	Measure	W=0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
4	4	50	Estrella	-0.08	-0.05	0.05	0.18	0.33	0.46	0.57	0.65	0.72	0.77	0.81
			BIC	3.00	2.97	2.87	2.72	2.54	2.36	2.19	2.03	1.89	1.75	1.64
			LR	4.04	5.87	10.95	18.23	27.15	36.35	45.00	52.84	59.86	66.84	72.40
4	4	250	Estrella	-0.02	0.02	0.13	0.27	0.42	0.54	0.64	0.72	0.78	0.83	0.87
			BIC	2.84	2.80	2.69	2.52	2.34	2.14	1.96	1.79	1.64	1.51	1.39
			LR	3.99	14.18	42.65	84.10	131.23	179.81	225.02	267.06	304.27	337.46	367.51
7	4	50	Estrella	-0.08	-0.02	0.13	0.33	0.51	0.65	0.76	0.82	0.88	0.91	0.94
			BIC	4.12	4.06	3.89	3.65	3.37	3.10	2.82	2.60	2.38	2.18	2.02
			LR	4.02	7.00	15.49	27.68	41.81	55.25	68.99	80.13	91.09	101.15	109.00
7	4	250	Estrella	-0.02	0.04	0.18	0.37	0.54	0.67	0.77	0.84	0.89	0.92	0.94
			BIC	3.96	3.91	3.75	3.52	3.24	2.97	2.71	2.48	2.26	2.08	1.90
			LR	3.97	17.57	57.29	116.14	183.99	252.48	316.61	376.06	429.78	475.73	519.04
10	4	50	Estrella	-0.08	-0.04	0.07	0.21	0.38	0.52	0.64	0.74	0.81	0.87	0.90
			BIC	4.84	4.80	4.68	4.51	4.29	4.06	3.80	3.55	3.31	3.08	2.90
			LR	4.05	6.01	11.94	20.16	31.52	42.96	55.78	68.24	80.49	91.89	100.97
10	4	250	Estrella	-0.02	0.03	0.15	0.31	0.47	0.62	0.74	0.82	0.88	0.92	0.94
			BIC	4.68	4.63	4.51	4.31	4.06	3.78	3.49	3.22	2.96	2.72	2.51
			LR	4.04	15.06	46.87	96.67	159.19	228.72	301.39	368.70	434.50	492.35	545.39
4	4	50	Estrella	-0.08	-0.04	0.08	0.27	0.48	0.67	0.80	0.89	0.94	0.97	0.99
One dominant			BIC	3.00	2.96	3.00	2.96	2.83	2.62	2.33	1.99	1.67	1.37	1.10
			LR	4.04	6.12	12.57	23.18	37.72	54.53	70.81	85.97	99.31	111.75	121.08
4	4	250	Estrella	-0.02	0.02	0.13	0.29	0.47	0.62	0.75	0.84	0.90	0.94	0.97
One dominant			BIC	2.84	2.81	2.69	2.50	2.26	2.00	1.74	1.48	1.25	1.03	0.84
			LR	3.99	13.79	43.28	89.96	149.36	215.23	281.28	344.23	402.82	456.70	506.46
7	4	50	Estrella	-0.08	-0.02	0.15	0.36	0.57	0.74	0.86	0.93	0.97	0.99	1.00
One dominant			BIC	4.12	4.06	3.88	3.61	3.26	2.88	2.45	2.06	1.69	1.35	1.07
			LR	4.03	7.12	16.10	29.87	47.33	66.42	87.80	107.40	125.97	142.86	156.80
7	4	250	Estrella	-0.02	0.04	0.19	0.39	0.59	0.75	0.86	0.94	0.97	0.99	1.00
One dominant			BIC	3.96	3.91	3.74	3.48	3.15	2.78	2.38	1.97	1.60	1.25	0.97
			LR	4.01	17.93	59.37	124.65	207.89	301.18	401.23	501.96	595.37	682.03	753.21
10	4	50	Estrella	-0.08	-0.04	0.08	0.24	0.45	0.63	0.79	0.89	0.96	0.99	1.00
One dominant			BIC	4.84	4.80	4.67	4.48	4.18	3.83	3.41	2.92	2.41	1.90	1.43
			LR	4.06	6.05	12.46	22.13	36.89	54.36	75.42	99.79	125.17	150.94	174.49
10	4	250	Estrella	-0.02	0.03	0.15	0.33	0.52	0.69	0.83	0.92	0.97	0.99	1.00
One dominant			BIC	4.68	4.63	4.50	4.28	3.98	3.61	3.19	2.73	2.24	1.78	1.35
			LR	4.06	15.42	48.34	103.66	177.75	269.96	376.24	491.82	613.26	729.42	835.68
4	4	50	Estrella	-0.08	-0.04	0.07	0.22	0.38	0.53	0.65	0.73	0.80	0.84	0.88
Two dominant			BIC	3.00	2.97	2.85	2.68	2.47	2.24	2.03	1.85	1.69	1.55	1.43
			LR	4.04	5.96	11.65	20.12	30.88	42.21	52.69	61.53	70.00	76.88	82.73
4	4	250	Estrella	-0.02	0.03	0.14	0.31	0.47	0.61	0.72	0.80	0.86	0.90	0.93
Two dominant			BIC	2.84	2.80	2.68	2.49	2.26	2.02	1.80	1.60	1.42	1.26	1.12
			LR	3.99	14.67	46.11	93.93	150.41	210.12	265.21	315.68	360.22	399.87	433.99
7	4	50	Estrella	-0.08	-0.02	0.15	0.38	0.59	0.74	0.84	0.91	0.95	0.97	0.98
Two dominant			BIC	4.12	4.06	3.88	3.58	3.23	2.88	2.53	2.23	1.95	1.71	1.53
			LR	4.02	7.18	16.47	31.12	48.60	66.34	83.85	98.91	112.90	124.78	133.74
7	4	250	Estrella	-0.02	0.04	0.20	0.41	0.61	0.75	0.85	0.92	0.95	0.97	0.98
Two dominant			BIC	3.96	3.91	3.73	3.45	3.11	2.76	2.42	2.10	1.82	1.58	1.37
			LR	3.98	18.15	62.39	132.64	216.66	304.76	390.13	470.52	539.00	600.49	652.09
10	4	50	Estrella	-0.08	-0.04	0.07	0.21	0.38	0.52	0.64	0.74	0.81	0.87	0.90
Two dominant			BIC	4.84	4.80	4.68	4.51	4.29	4.06	3.80	3.55	3.31	3.08	2.90
			LR	4.05	6.01	11.94	20.16	31.52	42.96	55.78	68.24	80.49	91.89	100.97
10	4	250	Estrella	-0.02	0.03	0.15	0.31	0.47	0.62	0.74	0.82	0.88	0.92	0.94
Two dominant			BIC	4.68	4.63	4.51	4.31	4.06	3.78	3.49	3.22	2.96	2.72	2.51
			LR	4.04	15.06	46.87	96.67	159.19	228.72	301.39	368.70	434.50	492.35	545.39

[18] The models are estimated without a check for values of Estrella-R^2.

Shrinkage estimators start converging to the MLE for values of Estrella-R^2 less than 20% when N=50 and close to 12% when N=250. The risk gain is smaller compared to the case of collinearity between variables, and similar to the orthonormal case. Increasing the number of alternatives does not improve the performance of the shrinkage estimators. The results do not change if one or more alternatives are dominant. The only important difference we observed is a significant decrease in absolute risk when one of the alternatives is dominant and we increase the number of alternatives in small samples. We still do not have an explanation about the improved estimator performance in this particular case. In terms of individual performance of the estimators, generally more shrinkage offers larger risk improvement, but there is no dominant shrinkage estimator over the entire parameter space.

- Weighted Squared Error Loss

The results are very similar to the orthonormal case, and there are no important differences between the models where all alternatives are equally likely, or where one or more of the alternatives are dominant, with the exception of one dominant alternative when N=50. In the latter case, comparable to the case of squared error loss, there is a significant reduction in absolute risk as the number of alternatives increases. Again, in terms of relative risk, Stein rule estimators dominate the MLE for the entire parameter space, and convergence occurs for values of Estrella-R^2 less than 20%, as in the case of squared error loss.

- Out-Of-Sample Prediction Loss

In the case of severe collinearity we found that the results are similar to the orthonormal case, and as expected, this result was confirmed also in models with

collinearity between alternative-specific variables. The results for relative risk, as in the previous cases, show that shrinkage improves prediction, but there is no dominant shrinkage estimator over the entire parameter space.

Table 3.13: Count of samples excluded from the Monte Carlo experiment[19]

J	K	N	Equal shares	One dominant	Half dominant	W=.6	W=.7	W=.8	W=.9	W=1
Severe Collinearity Between Two Variables										
4	4	50	X							0
4	4	250	X							0
4	4	50		X				4	100	762
4	4	250		X						317
4	4	50			X			1	2	7
4	4	250			X					0
4	7	50	X							0
4	7	250	X							0
4	7	50		X			4	60	386	1465
4	7	250		X					20	1311
4	7	50			X		1	8	25	96
4	7	250			X					5
4	10	50	X						2	3
4	10	250	X							0
4	10	50		X		407	3376	24874	256328	4292121
4	10	250		X				80	4582	663201
4	10	50			X	2	25	64	177	377
4	10	250			X			1	155	1571
Severe Collinearity Among All Variables										
4	4	50	X							0
4	4	250	X							0
4	4	50		X				3	29	143
4	4	250		X						192
4	4	50			X				1	33
4	4	250			X					1
4	7	50	X			23	99	268	581	1123
4	7	250	X					5	47	355
4	7	50		X		90	323	684	1479	3512
4	7	250		X			15	684	9546	240584
4	7	50			X	52	203	483	1058	1845
4	7	250			X			77	1162	10594
4	10	50	X			353	789	1630	3043	5462
4	10	250	X			233	1752	10794	77817	845548
4	10	50		X		368	1109	2638	5355	8713
4	10	250		X		-	-	-	-	-
4	10	50			X	408	1234	2799	6484	13173
4	10	250			X	-	-	-	-	-

[19] The models with N=250, K=10, and one or two dominant alternatives, were estimated without a check for values of Estrella-R^2 because of the high computation time.

3.5 Conclusions

The objective of this study is to explore the properties of Stein-rule estimators when collinearity is present among the explanatory variables. We look at collinearity between the explanatory variables within an alternative, and between alternatives. We analyze three different types of models: where all alternatives are equally likely, where one alternative is dominant, and where half of the alternatives are dominant.

Our results show that in conditional logit the results are going to be influenced not only by the degree of collinearity, but also by the type of collinearity. We found that when collinearity is present between alternative-specific variables, its effects disappear in the estimation process. In this case the results for relative and absolute risk are comparable to the orthonormal model: Stein rule estimators have lower risk than the maximum likelihood estimator both in terms of estimation and in terms of prediction for the entire parameter space under consideration, and the existence of collinearity between alternatives does not affect the results. When collinearity is present between the variables within alternatives, the risk improvement is larger compared to the orthonormal model, and increases with the degree of collinearity and with the number restrictions in the model.

In small samples, when the number of variables increases and severe collinearity is present among all variables, we found a very significant increase in risk of all estimators. Although shrinkage offers risk improvement, such complex models should not be estimated with small number of observations when severe collinearity is present among the regressors. In terms of out-of-sample prediction we did not find significant differences between the different types or degree of collinearity.

4 Applications of Shrinkage Estimation in Multinomial Choice Models

4.1 Introduction

The previous two chapters studied the risk properties of Stein rule estimators in the context of the conditional logit model, where the explanatory variables were simulated in a Monte Carlo experiment. This chapter extends our analysis to several real applications of the conditional and multinomial logit models. The analysis is performed on three economic and marketing data sets. We are interested in determining whether shrinkage can improve out-of-sample prediction for models with different numbers of parameters and different quality of non-sample information.

This chapter is organized as follows. In section 2 we describe the way in which our analysis is conducted. Section 3 estimates a conditional logit model about the choice between four brands of saltine crackers. Section 4 estimates a conditional logit model about the choice between seven brands of soft drink beverages. Section 5 estimates a multinomial logit model about car ownership, and section 6 concludes.

4.2 Estimation Method

Each model is estimated by maximum likelihood and Stein-rule estimation procedures. The Stein rule estimator for the conditional logit is given by

$$\delta = \left(1 - \frac{c}{u}\right)\beta_U + \left(\frac{c}{u}\right)\beta_R, \tag{4.1}$$

and the corresponding positive-rule is given by

$$\delta^+ = \left[1 - \frac{c}{u}\right]_+ \beta_U + \left(\frac{c}{u}\right)\beta_R, \tag{4.2}$$

where β_U and β_R are the unrestricted and the restricted maximum likelihood estimates respectively. The way in which shrinkage estimation works, and the choice of a test-statistic *u* and a shrinkage constant *c* are discussed in detail in the orthonormal case.

The estimation of each model is performed using the econometrics software Limdep. The code for the Stein-rule estimator and the mean square error of prediction can be provided upon request. We switched to Limdep in order to show that shrinkage estimation is not only easy to understand, but also easy to do using standard econometrics software. The analysis is done in the following steps:

- extract a certain number of observations from the original data to create a holdout sample for out-of-sample prediction;
- select the restricted models to be used in the Stein rule estimation;
- estimate the model coefficients and evaluate out-of-sample predictive ability for each estimator.

4.3 Saltine Crackers Data

4.3.1 Data

The data set is a scanner panel of 136 households in Georgia observed for two years. Each household has a choice of four brands of saltine crackers: Nabisco, Sunshine, Keebler, and a collection of private labels. The explanatory variables are the actual price of the purchased brand and the shelf prices of other brands, and three dummy variables indicating whether the brand was on display, featured in a newspaper, or jointly on display and featured at the time of purchase. There are a total of 3292 observations. To allow for out-of-sample prediction, the last purchase of each household was used to

create a holdout sample of 136 observations. The remaining 3156 observations were used for parameter estimation. The data set was originally provided by Information Resources, Inc. and used in the estimation of a conditional logit model by Frances and Paap (2001).

Descriptive information about the data is presented in Tables 4.1 and 4.2. Keebler is the most expensive brand with average price of $1.08. The Private label has the lowest average price of $0.68. There is no collinearity between the explanatory variables.

Table 4.1: Saltine Data - Correlations Between the Explanatory Variables

Correlations:	Price	Display	Feature	Display & Feature
Price	1	0.04	-0.03	-0.13
Display	0.04	1	-0.06	-0.07
Feature	-0.03	-0.06	1	-0.03
Display & Feature	-0.13	-0.07	-0.03	1

Table 4.2: Saltine Data – Relative Shares

Brand	Share
Private Label	31.4%
Sunshine	7.3%
Keebler	6.9%
Nabisco	54.4%

4.3.2 Estimation

Table 4.3 shows the maximum likelihood parameter estimates and their corresponding standard errors. The model contains three alternative specific dummy variables, with Nabisco omitted to serve as the reference group. The alternative specific parameters for the Private label, Sunshine, and Keebler are negative, indicating that Nabisco is the market leader. This result is also confirmed by the relative shares of each

brand: 54% of individuals choose Nabisco, followed by the Private label (32%), Sunshine (7.4%), and Keebler (6.8%).

Table 4.3: Saltine data, maximum likelihood estimates

Variables	Parameter	Standard error
Intercepts		
Private label	-1.814*	0.104
Sunshine	-2.465*	0.082
Keebler	-1.968*	0.075
Marketing variables		
Price	-3.172*	0.216
Display	0.049	0.068
Feature	0.412*	0.151
Display and feature	0.580*	0.119
Max log-likelihood value	-3215.83	

Note: *Significant at the 0.01 level, N=3156.

The negative price coefficient agrees with economic theory, indicating that an increase in the price of a brand decreases the probability of that brand being purchased. The positive coefficients of the promotional variables show that the probability of choosing a brand is higher if that brand was on display or featured at the time of purchase. However, the parameter of a single display is not statistically significant. All other parameters are significant at the 0.01 level.

In order to introduce different sets of non-sample information, we modified the original model by creating alternative-specific variables, thus allowing the parameters of the model to vary across brands. This is not unreasonable to do, because having constant price effects is only an assumption. The full model is presented in Table 4.4.

Table 4.4: Saltine data, full model

Variables	MLE	Standard error
Intercepts		
Private label	-3.031*	0.366
Sunshine	-1.339**	0.609
Keebler	0.438	0.811
Marketing variables		
Price Private	-1.563*	0.345
Price Sunshine	-4.689*	0.578
Price Keebler	-5.577*	0.720
Price Nabisco	-3.347*	0.282
Display Private	-0.495*	0.179
Display Sunshine	0.170	0.210
Display Keebler	0.244	0.230
Display Nabisco	0.040	0.083
Feature Private	-0.045	0.359
Feature Sunshine	0.351	0.395
Feature Keebler	0.544	0.401
Feature Nabisco	0.430**	0.204
Display and feature Private	0.246	0.210
Display and feature Sunshine	0.959*	0.314
Display and feature Keebler	0.615**	0.288
Display and feature Nabisco	0.650*	0.196
Max log-likelihood value	-3177.54	

Note: *Significant at the 0.01 level, **Significant at the 0.05 level, N=3156.

The model contains three alternative specific dummy variables, with Nabisco omitted to serve as the reference group. The alternative specific parameters for the Private label, Sunshine, and Keebler are negative, indicating that Nabisco is the market leader. This result is also confirmed by the relative shares of each brand: 54% of individuals choose Nabisco, followed by the Private label (32%), Sunshine (7.4%), and Keebler (6.8%). For the Stein-rule estimation, five restricted models were chosen, namely:

Restriction Set 1. All alternative-specific parameters, including the intercepts, are equal.

Restriction Set 2. All alternative-specific parameters, except for the intercepts, are equal. In this case the restricted model is the original model, which assumes that the model parameters do not vary across alternatives.

Restriction Set 3. All alternative-specific parameters, except for the intercepts, are equal and the parameters for display and display&feature equal zero. We assume that feature is the only promotional variable, which influences the choice of an alternative.

Restriction Set 4. The price coefficients for the three national brands are equal, and the parameters for display equal zero. We assume that the price coefficients would differ between a national brand and a private label, but there is no difference between the national brands. In addition, we assume that display alone does not influence the choice of an alternative, but we include the interaction variable display&feature.

Restriction Set 5. The parameters for Sunshine and Keebler are equal for each explanatory variable. The descriptive statistics and the relative shares of the four brands show that there is no significant difference between Sunshine and Keebler therefore we assume a single parameter for the two brands.

The estimated coefficients obtained by the Stein-rule estimator for each restriction are reported in Table 4.5.

Table 4.5: Saltine data, Stein-rule and MLE estimates

Variables	MLE	Stein R1	Stein R2	Stein R3	Stein R4	Stein R5
Intercepts						
Private label	-3.031*	-3.011	-2.872	-2.840	-3.171	-2.808
Sunshine	-1.339**	-1.330	-1.486	-1.515	-1.775	-1.426
Keebler	0.438	0.436	0.124	0.061	-0.366	0.027
Marketing variables						
Price Private	-1.563*	-1.559	-1.773	-1.815	-1.569	-1.857
Price Sunshine	-4.689*	-4.664	-4.491	-4.451	-4.380	-4.412
Price Keebler	-5.577*	-5.547	-5.263	-5.201	-4.976	-5.138
Price Nabisco	-3.347*	-3.331	-3.324	-3.319	-3.481	-3.315
Display Private	-0.495*	-0.487	-0.424	-0.418	-0.332	-0.405
Display Sunshine	0.170	0.174	0.155	0.144	0.114	0.139
Display Keebler	0.244	0.247	0.218	0.206	0.163	0.199
Display Nabisco	0.040	0.045	0.041	0.034	0.027	0.033
Feature Private	-0.045	-0.038	0.015	0.027	-0.036	0.039
Feature Sunshine	0.351	0.355	0.359	0.361	0.431	0.362
Feature Keebler	0.544	0.547	0.527	0.523	0.570	0.520
Feature Nabisco	0.430**	0.434	0.428	0.428	0.435	0.427
Display and feature Private	0.246	0.251	0.289	0.207	0.249	0.201
Display and feature Sunshine	0.959*	0.959	0.909	0.808	1.034	0.783
Display and feature Keebler	0.615**	0.618	0.610	0.519	0.668	0.502
Display and feature Nabisco	0.650*	0.653	0.641	0.548	0.642	0.531
Max log-likelihood value	-3177.5	-4165.2	-3215.8	-3215.8	-3183.6	-3178.5

Table 4.6 shows the mean squared error of out-of-sample prediction for the MLE, Stein-rule and Restricted estimators. The shaded values indicate the best predictor.

Table 4.6: Saltine data, MSE of out-of-sample prediction

Restriction:	MLE	Stein	Restricted	LR statistic	# of restrictions
1	0.142369	0.142275	0.165324	1975.23	15
2	0.142369	0.141315	0.135790	76.59	12
3	0.142369	0.141555	0.138451	76.59	14
4	0.142369	0.142153	0.142118	12.12	6
5	0.142369	0.142367	0.142367	1.90	6

The values of the likelihood ratio test statistic and the number of restrictions determine the degree of shrinkage towards the restricted model. In the first three models the *LR* statistic is relatively large which means that the restrictions would be rejected. For restriction sets 1 and 2 it implies that a model assuming constant price effects is misspecified. In terms of shrinkage estimation it implies that larger weight is placed on the unrestricted model. Restriction 5 results in a *LR* value smaller than the shrinkage constant, in which case the positive-part Stein-rule takes the values of the restricted model in order to prevent over-shrinkage and to preserve the signs of the estimated coefficients. Overall, the values of the estimates are very similar, and although the Stein-rule performs better than the MLE, the improvement is not significant. The MLE performs very well because of the large number of observations and good quality of data, which implies that shrinkage cannot lead to significant improvement in estimation and prediction.

4.4 Cola Data

4.4.1 Data

The data set is a scanner panel, kindly provided by Ron Niedrich and Danny Weathers from the marketing department at LSU. The data are collected from five stores over a 104-week period. There are 287 households, making purchases on a total of 3,546 occasions. Each household has a choice of seven brands of two-liter carbonated beverages. The brands are: Pepsi, 7-UP, Coca-Cola Classic, Diet Coke, Diet Rite, Diet Pepsi, and Diet 7-UP. The explanatory variables are the shelf prices of each brand, two dummy variables indicating whether the brand was on display or featured at the time of purchase, and a variable measuring brand loyalty, created by Niedrich et al. (2004)

following the marketing literature. To allow for out-of-sample prediction, the last purchase of each household was used to create a holdout sample of 287 observations. The remaining 3,259 observations were used for parameter estimation.

Descriptive information about the data is presented in Tables 4.7 and 4.8.

Table 4.7: Cola Data - Correlations Between the Explanatory Variables

Correlations:	Price	Display	Feature	Loyalty
Price	1	-0.61	-0.64	-0.06
Display	-0.61	1	0.56	0.05
Feature	-0.64	0.56	1	0.04
Loyalty	-0.06	0.05	0.04	1

Table 4.8: Cola Data – Relative Shares

Brand	Share
Pepsi	17.7%
7UP	18.5%
Coke	14.4%
Diet Coke	13.8%
Diet Rite	12.5%
Diet Pepsi	10.7%
Diet 7UP	12.3%

There are no significant differences in prices among the seven brands. The overall average price is $1.2. The most expensive brands are Pepsi and Diet Pepsi, with an average price of $1.26, closely followed by Coke and Diet Coke ($1.23). The least expensive brand is Diet Rite with an average price of $1.15. Similarly, the loyalty variable does not change much between brands. Interestingly, consumers on average appear most loyal to Pepsi and 7-UP, and least loyal to Diet Pepsi and Diet 7-UP. There is moderate negative correlation between price and the promotional variables, which suggests that a brand might be on display or featured when the product is on sale. There is a moderate positive correlation between feature and display, indicating that the two

promotional techniques are usually used together. Loyalty is not correlated with prices and promotion.

4.4.2 Estimation

Table 4.9 shows the maximum likelihood parameter estimates and their corresponding standard errors. The model contains six alternative specific dummy variables, with Diet 7-UP omitted to serve as the reference group. Except for Diet Rite, all alternative specific parameters are positive, indicating that the brands are preferred compared to the reference brand. The parameter for Diet Rite and Diet Pepsi are not statistically significant.

Table 4.9: Cola data, maximum likelihood estimates

Variables	Parameter	Standard error
Intercepts		
Pepsi	0.432*	0.082
7-UP	0.191*	0.078
Coke	0.313*	0.083
Diet Coke	0.216*	0.083
Diet Rite	-0.114	0.088
Diet Pepsi	0.130	0.088
Marketing variables		
Price	-1.862*	0.146
Display	0.586*	0.072
Feature	-0.037	0.070
Loyalty	3.290*	0.061
Max log-likelihood value	-4153.40	

Note: *Significant at the 0.01 level, N=3259.

The relative shares show that 18.5% of individuals choose 7-UP, followed by Pepsi (17.7%). Diet Pepsi has the lowest share (10.7%), followed by Diet 7-UP and Diet

Rite. The parameters of the marketing variables have the expected signs, with the exception of feature, which is not statistically significant. All other parameters are significant at the 0.01 level.

In order to introduce different sets of non-sample information, we modified the original model by creating alternative-specific variables, thus allowing the parameters of the model to vary across brands. The full model has 28 variables and the estimates are presented in Table 4.2.2 in the appendix. For the Stein-rule estimation, five restricted models were chosen, namely:

Restriction Set 1. All alternative-specific parameters, including the intercepts, are equal.

Restriction Set 2. All alternative-specific parameters, except for the intercepts, are equal.

Restriction Set 3. The alternative-specific parameters and the intercepts are equal for: Pepsi and 7-UP, and Coke and Diet Coke.

Table 4.10 shows the mean squared error of out-of-sample prediction for the MLE, Stein-rule and Restricted estimators. The shaded values indicate the best predictor.

Table 4.10: Cola data, MSE of out-of-sample prediction

Restriction:	MLE	Stein	Restricted	LR statistic	# of restrictions
1	0.699	0.698	0.697	150.12	25
2	0.699	0.699	0.703	94.27	19
3	0.699	0.697	0.691	48.99	10

The values of the likelihood ratio test statistic and the number of restrictions determine the degree of shrinkage towards the restricted model. In all three models the *LR* statistic is relatively large and larger weight is placed on the unrestricted model therefore we do not expect significant difference between the Stein-rule and the MLE.

The estimated coefficients obtained by the Stein-rule estimator for each restriction are reported in Table 4.11.

Table 4.11: Cola data, MSE of out-of-sample prediction

Variables	MLE	Stein R1	Stein R2	Stein R3
Intercepts				
Pepsi	0.856	0.725	0.779	0.880
7-UP	0.867**	0.734	0.745	0.889
Coke	-0.009	-0.008	0.048	0.014
Diet Coke	-0.234	-0.198	-0.154	-0.175
Diet Rite	-1.004**	-0.850	-0.842	-0.840
Diet Pepsi	0.168	0.142	0.160	0.141
Marketing variables				
Feature	-0.086	-0.073	-0.071	-0.072
Price Pepsi	-2.405*	-2.288	-2.300	-2.358
Price 7-UP	-2.263*	-2.168	-2.184	-2.239
Price Coke	-1.638*	-1.638	-1.671	-1.616
Price Diet Coke	-1.648*	-1.646	-1.679	-1.624
Price Diet Rite	-1.595*	-1.602	-1.636	-1.665
Price Diet Pepsi	-1.956*	-1.907	-1.932	-1.886
Price Diet 7-UP	-2.054*	-1.990	-2.012	-2.004
Display Pepsi	0.860*	0.823	0.809	0.798
Display 7-UP	0.222	0.283	0.286	0.264
Display Coke	0.496*	0.515	0.511	0.493
Display Diet Coke	0.536*	0.548	0.543	0.527
Display Diet Rite	0.996*	0.938	0.920	0.953
Display Diet Pepsi	0.727*	0.710	0.700	0.731
Display Diet 7-UP	0.465*	0.488	0.485	0.485
Loyalty Pepsi	3.161*	3.183	3.184	3.115
Loyalty 7-UP	2.592*	2.701	2.718	2.639
Loyalty Coke	3.089*	3.122	3.125	3.123
Loyalty Diet Coke	3.556*	3.517	3.508	3.514
Loyalty Diet Rite	4.334*	4.176	4.146	4.323
Loyalty Diet Pepsi	2.890*	2.953	2.962	2.895
Loyalty Diet 7-UP	3.832*	3.751	3.734	3.831
Max log-likelihood value	-4106.4	-4181.5	-4153.5	-4130.9

Note: *Significant at the 0.01 level, **Significant at the 0.05 level, N=3259.

Overall, the values of the estimates are very similar, and the performance of the Stein-rule is either better or equal to the performance of the MLE, but there is no significant improvement. The choice of restrictions on this model is not exhaustive, and we are going to try other choices of non-sample information.

4.5 Car Ownership Data

4.5.1 Data

The data set is on private car ownership of Dutch households in 1989, made available and used for estimation of different choice models by Cramer (2003). The data set consists of 2820 records, one for each household, who choose between four categories of private car ownership, namely: None, Used, New, and More. The choice is assumed to depend on the following individual-specific variables: *Income* per equivalent adult, measured in Dutch guilders per annum; *Size*, the size of the household, measured per equivalent adults (1 for the first adult, 0.7 for other adults, and 0.5 for children); *Age*, the age of the head of household, measured by five year classes, starting with the class "below 20"; *Urba*, the degree of urbanization, measured on a six-point scale from countryside (1) to city (6); and *Buscar*, a (0,1) dummy variable for the presence of a business car in the household. To allow for out-of-sample prediction, the last 400 observations were used as a holdout sample. The remaining 2,420 observations were used for parameter estimation.

Descriptive information and histograms of the data are presented in Tables 4.12 and 4.13. There is no correlation between the explanatory variables.

Table 4.12: Car Ownership Data - Correlations Between the Explanatory Variables

Correlations:	Income	Household Size	Age	Urbanization	Business Car
Income	1	-0.14	-0.09	-0.14	-0.14
Household Size	-0.14	1	-0.15	-0.08	-0.14
Age	-0.09	-0.15	1	-0.15	-0.08
Urbanization	-0.14	-0.08	-0.15	1	-0.14
Business Car	-0.14	-0.14	-0.08	-0.14	1

Table 4.13: Car Ownership Data – Relative Shares

Choice	Share
None	35.8%
Used	33.5%
New	24.5%
More	6.2%

4.5.2 Estimation

Table 4.14 shows the maximum likelihood parameter estimates and their corresponding standard errors. The alternative specific dummy variables are positive and statistically significant. The reference category is *More*, which appears as the least preferred choice of car ownership. This is also supported by the relative shares of each choice: 36% of the individuals in the sample do not own a car, 34% own a used car, 25% own a new car, and only 6% own more than one car.

The coefficients for *Urban* are not significant for each alternative, *Buscar* is not significant if the choice is new or used car, and *Age* is not significant if the choice is not to own a private car. All other parameters are significant at the 0.01 level.

For the Stein-rule estimation, four restricted models were chosen, namely:

Restriction Set 1. Urbanization does not affect the choice of car ownership.

Restriction Set 2. Urbanization and the presence of a business car do not affect the choice.

Restriction Set 3. Urbanization, business car and age do not affect the choice.

Restriction Set 4. Urbanization, business car, age and household size do not affect the choice.

Table 4.14: Car Ownership data, maximum likelihood estimates

Variables	Parameter	Standard error
None	49.415*	3.038
None x Income	-10.403*	0.670
None x HH size	-13.454*	0.755
None x Age	0.044	0.037
None x Urban	0.080	0.055
None x Bus car	3.598*	0.369
Used	33.344*	2.862
Used x Income	-6.512*	0.628
Used x HH size	-7.570*	0.714
Used x Age	-0.151*	0.037
Used x Urban	-0.049	0.052
Used x Bus car	0.595	0.379
New	19.844*	2.815
New x Income	-3.613*	0.617
New x HH size	-7.269*	0.713
New x Age	-0.014	0.036
New x Urban	-0.034	0.053
New x Bus car	0.574	0.383
Max log-likelihood value	-2874.9	

Note: *Significant at the 0.01 level, N=2420.

The estimated coefficients obtained by the Stein-rule estimator for each restriction are reported in Table 4.15. Table 4.16 shows the mean squared error of out-of-sample prediction for the MLE, Stein-rule and Restricted estimators. The shaded values indicate the best predictor.

Table 4.15: Car Ownership Data, MLE and Stein-Rule Estimates

Variables	MLE	Stein R1	Stein R2	Stein R3	Stein R4
None	49.415*	49.042	49.030	48.942	48.928
None x Income	-10.403*	-10.319	-10.316	-10.295	-10.297
None x HH size	-13.454*	-13.395	-13.386	-13.374	-13.321
None x Age	0.044	0.045	0.044	0.044	0.044
None x Urban	0.080	0.079	0.079	0.079	0.079
None x Buss car	3.598*	3.590	3.570	3.564	3.563
Used	33.344*	33.093	33.085	33.025	33.016
Used x Income	-6.512*	-6.457	-6.455	-6.444	-6.444
Used x HH size	-7.570*	-7.527	-7.526	-7.517	-7.496
Used x Age	-0.151*	-0.151	-0.151	-0.150	-0.150
Used x Urban	-0.049	-0.049	-0.049	-0.049	-0.049
Used x Buss car	0.595	0.590	0.590	0.589	0.589
New	19.844*	19.694	19.690	19.654	19.649
New x Income	-3.613*	-3.580	-3.579	-3.572	-3.574
New x HH size	-7.269*	-7.242	-7.240	-7.235	-7.197
New x Age	-0.014	-0.014	-0.014	-0.014	-0.014
New x Urban	-0.034	-0.033	-0.033	-0.033	-0.033
New x Buss car	0.574	0.573	0.570	0.569	0.569
Max log-likelihood value	-2874.9	-3139.9	-3324.5	-3397.4	-3535.1

Note: *Significant at the 0.01 level, **Significant at the 0.05 level, N=2420.

Table 4.16: Car ownership data, MSE of out-of-sample prediction

Restriction:	MLE	Stein	Restricted	LR statistic	# of restrictions
1	0.5457	0.5457	0.5879	530.1	6
2	0.5457	0.5457	0.6468	899.4	9
3	0.5457	0.5458	0.6660	1045.0	12
4	0.5457	0.5459	0.7012	1320.6	15

The values of the likelihood ratio test statistic in all models are very large, which implies a very small degree of shrinkage towards the restricted model. There are virtually no differences between the MLE and the Stein estimators, and the restricted estimator has the worst performance in all cases. In this model shrinkage does not

improve prediction, which is likely to be the result of good performance of MLE. Moreover, the restricted models were chosen for illustrative purposes only and not backed by theory, which resulted in a poor quality of non-sample information.

4.6 Conclusions

We estimated three choice models which differed by the number of alternatives, the number of variables, and the quality of non-sample information, introduced in the form of restrictions on the parameter coefficients. The performance of the MLE and the shrinkage estimators was very similar and the values of the estimated coefficients and the estimated mean squared errors of prediction were very close numerically. Stein rule did not show significant improvement over MLE because none of the conditions under which shrinkage offers significant gain was met. The models were estimated with close to perfect data with sufficiently large number of observations to assure very good performance of the maximum likelihood estimator. In addition, the non-sample information was chosen for illustrative purposes and we have no reason to believe that the imposed restrictions were correct. Even in these conditions, consistent with our previous results, the Stein rule did not perform worse, and in some cases performed slightly better than the MLE. Shrinkage offers risk gains in repeated samples, and marketing researchers often estimate the same model using different data for different products or different geographical markets. Therefore, we expect overall larger gains from shrinkage applied to marketing data. In addition, when imposing restrictions, we can significantly improve the quality of non-sample information by collaborating with a product specialist, thus creating conditions for the Stein-rule estimator to offer significant risk improvement compared to maximum likelihood estimation.

5 Conclusions

The objective of this study is to compare the performance of a Stein-like estimator to the MLE in the context of the conditional logit model. We explored the risk properties of the estimators via a Monte Carlo experiment for the orthonormal model as well as when collinearity is present among the regressors. Finally, we applied shrinkage estimation to three data sets to evaluate the performance of the estimators in terms of out-of-sample prediction.

In the Monte Carlo experiment we looked at three different types of models: a model where all alternatives are equally likely, a model where one alternative is dominant, and a model where half of the alternatives are dominant.

In the orthonormal case, when the number of alternatives equals to four, the results of the Monte Carlo experiment showed that Stein rule estimators have lower risk than the maximum likelihood estimators both in terms of estimation and in terms of prediction for the entire parameter space under consideration. The risk improvement is larger in small samples and for small degrees of specification error. In addition, Stein rule performance improves relative to the MLE when the number of variables increases. In large samples the performance of all estimators improves because both the bias and the variance of the estimators decrease with an increase in the number of observations. The results for relative risk showed no significant difference between the three cases.

When we increase the number of alternatives we confirm the result that shrinkage leads to risk improvement over the entire parameter space. Higher number of alternatives does not favor shrinkage in estimation but the performance of all estimators improves. In

out-of-sample prediction, higher number of alternatives expands the parameters space for which Stein rule estimators offer risk improvement over MLE in small samples.

Our next objective is to explore the properties of Stein-rule estimators when collinearity is present among the explanatory variables. We look at collinearity between the explanatory variables within an alternative, and also at collinearity between specific variables between alternatives. We analyze three different types of models: a model where all alternatives are equally likely, a model where one alternative is dominant, and a model where half of the alternatives are dominant.

Our results show that Stein rule estimators have lower risk than the maximum likelihood estimators both in terms of estimation and in terms of prediction for the entire parameter space under consideration. The risk improvement is larger compared to the orthonormal model, and increases with the degree of collinearity and with the number restrictions in the model. In most cases there is no dominant shrinkage estimator over the entire parameter space, although generally more shrinkage corresponds to larger risk improvement.

In small samples, when the number of variables increases and severe collinearity is present among all variables, we found out a very significant increase in risk of all estimators. Although shrinkage offers risk improvement, such complex models should not be estimated with small number of observations when severe collinearity is present among the regressors. In terms of out-of-sample prediction we did not find significant differences between the different kind or degree of collinearity. The best predictive performance of all estimators is in the model with one dominant alternative, followed by the model with two dominant alternatives.

When collinearity was present between alternative-specific variables, we found that it had no effect on estimation or prediction. This result raised some interesting questions about the effects of collinearity in conditional logit models, which we are going to pursue in the future.

The three models using real data showed that the performance of the MLE and the shrinkage estimators was very similar and the values of the estimated coefficients and the estimated mean squared errors of prediction were very close numerically. Stein rule did not show significant improvement over MLE because none of the conditions under which shrinkage offers significant gain was met. The models were estimated with close to perfect data with sufficiently large number of observations to assure very good performance of the maximum likelihood estimator. In addition, the non-sample information was chosen for illustrative purposes and we have no reason to believe that the imposed restrictions were correct. Even in these conditions, consistent with our previous results, the Stein rule did not perform worse, and in some cases performed slightly better than the MLE, which is what we observed in the Monte Carlo experiment when the signal-to-noise ratio increased and the shrinkage estimators converged to the MLE.

In terms of recommendations, Stein rule should be used instead of the MLE in small samples and when there is a high degree of collinearity among the regressors. It also offers a larger risk improvement in models with large number of variables, which allows for more restrictions and a higher degree of shrinkage. In addition, risk improvement is significant when the restrictions agree with the data. We can significantly improve the quality of non-sample information by collaborating with a

product specialist. If we are uncertain of the quality of prior information or we have a large number of observations, we can still use shrinkage estimation because it offers lower or equal risk relative to the maximum likelihood estimator over the entire space.

Our analysis can be further improved by exploring the effects of shrinkage using more real data and comparing the estimators' performance under different quality of non-sample information. We also consider introducing unobserved heterogeneity among individuals, which, when combined with Stein-rule estimation, may lead to significant risk improvement over MLE. For practical applications, we consider creating bootstrap confidence bands to allow the use of Stein-rule estimators also for interval estimation and hypothesis testing.

References

Adkins, L.C. and R.C. Hill (1989), "Risk Characteristics of a Stein-like Estimator for the Probit Regression Model," *Economics Letters*, 30, 19-26.

Adkins, L.C., Hill, R.C. and M. Kim (1993), "Shrinkage Estimation in Nonlinear Models," working paper, Louisiana State University, Baton Rouge.

Cramer, J.S. (2003), *Logit Models From Economics and Other Fields*, Cambridge University Press, Cambridge, UK.

Davidson, R. and J.G. MacKinnon (1984), "Convenient Specification Tests for Logit and Probit Models," *Journal of Econometrics*, 25, 241-262.

Davidson, R. and J.G. MacKinnon (2004), *Econometric Theory and Methods*, Oxford University Press, New York, NY.

Estrella, Arturo (1998), "A New Measure of Fit for Equations with Dichotomous Dependent Variables," *Journal of Business and Economics Statistics*, 16(2), 198-205.

Greene, W.H. (2003), *Econometric Analysis*, 5th edition, Prentice Hall, Englewood Cliffs, NJ.

Griffiths, W.E., R.C. Hill and G.G. Judge (1993), *Learning and Practicing Econometrics*, John Wiley and Sons, NY.

Griffiths, W.E., R.C. Hill and P.J. Pope (1987),"Small Sample Properties of Probit Model Estimators," *Journal of the American Statistical Association*, 399, 929-937.

Hill, R.C. (1987), "Modeling Multicollinearity and Extrapolation in Monte Carlo Experiments on Regression," *Advances in Econometrics*, 6, 127-155.

Hill, R.C., P.A. Cartwright and J.F. Arbaugh (1991), "The Use of Biased Predictors in Marketing Research," *International Journal of Forecasting*, 7, 271-282.

Judge, G.G., and M.E. Bock (1978), *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics*, North-Holland, Amsterdam.

Judge, G.G., R.C. Hill and M.E. Bock (1990), "An Adaptive Empirical Bayes Estimator of the Multivariate Normal Mean Under Quadratic Loss," *Journal of Econometrics*, 44, 189-213.

Judge, G.G., R.C. Hill, W. Griffiths and T.C. Lee (1988), *Theory and Practice of Econometrics*, John Wiley and Sons, New York, NY.

Judge, G.G., R.C. Hill, W. Griffiths, H. Lutkepohl and T.C. Lee (1988), *Introduction to the Theory and Practice of Econometrics*, 2nd edition, John Wiley and Sons, New York, NY.

Kamakura, W.A. and M. Wedel (2004), "An Empirical Bayes Procedure for Improving Individual-Level Estimates and Predictions From Finite Mixtures of Multinomial Logit Models," *Journal of Business and Economic Statistics*, 22, No.1, 121-125.

Kennedy, P. (1998), *A Guide to Econometrics*, 4th edition, MIT Press, Cambridge, MA.

Kim, M. and R.C. Hill (1995), "Shrinkage Estimation in Nonlinear Regression: The Box-Cox Transformation," *Journal of Econometrics*, 66, 1-33.

Knight, J.R., R.C. Hill and C.F. Sirmans (1993), "Stein Rule Estimation in Real Estate Appraisal," *The Appraisal Journal*, 539-544.

Train, K. (1986), *Qualitative Choice Analysis*, MIT Press, Cambridge, MA.

Train, K. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge, MA.

Vita

Vera Tabakova was born in Sofia, Bulgaria. She received her bachelor degree from The University of Sofia, Bulgaria, in 1993. In 1994 she received a Master of Arts Degree in economics from the Central European University in Prague, The Czech Republic. She returned to Bulgaria to work for the Ministry of Industry, and left again to pursue a Master of Business Administration Degree at Bocconi University in Milan, Italy, which she completed in 1996. In the fall semester of 1997 she became a graduate student at the Economics Department of Louisiana State University, where she taught principles of economics for three years. Vera will complete the degree of Doctor of Philosophy in May 2005.